DOCUMENT RESUME

ED 183 004 FL 010 934

AUTHOR	Spolsky, Bernard, Ed.
TITLE	Some Major Tests. Advances in Language Testing, Series: 1. Papers in Applied Linguistics.
INSTITUTION	Center for Applied Linguistics, Arlington, Va.
PUB DATE	79
NOTE	146p.
AVAILABLE FROM	Center for Applied Linguistics, 1611 North Kent Street, Arlington, VA 22209 ($5.50)
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	*Achievement Tests; Bilingual Education; Elementary Secondary Education; *English (Second Language); *French; Language Dominance; Language Instruction; *Language Proficiency; Language Research; *Language Tests; Postsecondary Education; Second Language Learning; *Speaking; Testing

ABSTRACT

This fascicle is intended to give some general notion of the state of the art in language testing by describing and reviewing several recent major tests. Its intention is to be representative rather than exhaustive. The following papers are included: (1) "Language Proficiency Tests Developed for the IEA International Study of Achievement in French as a Foreign Language," by John B. Carroll; (2) "The IEA International Study of Achievement Test of Proficiency in English as a Foreign Language," by E. Glyn Lewis; (3) "The Test of English as a Foreign Language (TOEFL)," by John W. Oller, Jr. and Bernard Spolsky; (4) "TOEFL in Recent Years," by Paul J. Angelis; (5) "The Oral Interview of the Foreign Service Institute," by Randall L. Jones; and (6) "A Review of the Bilingual Syntax Measure," by Ellen J. Rosansky. A bibliography completes the volume. (Author/AMH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED183004

Papers in Applied Linguistics

ADVANCES IN LANGUAGE TESTING SERIES: 1

Some Major Tests

edited by Bernard Spolsky

Bernard Spolsky, General Series Editor

FL010934

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CAL

G. Richard Tucker,
Dir

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Center for Applied Linguistics

June 1979
Copyright © 1979
by the Center for Applied Linguistics
1611 North Kent Street
Arlington, Virginia 22209

ISBN: 87281-074-7

Library of Congress Catalog Card Number: 78-60576

Printed in the United States of America

Preface

Language testing is one of the most obvious and important areas for activity in applied linguistics. It fits the general paradigm well: the problem is clearly language related, and the solution must come from linguistics and from another discipline--psychometrics--as well. Each of the major branches of linguistics--theoretical, psycholinguistic, and sociolinguistic--has its own special role to play, and each has exerted its influence on the development of the field. The involvement of psychometrics is a necessity as well. Good language testing needs to be based on relevant knowledge from applied linguistics and from psychometrics.

Given this, and considering the social relevance of the field, it is appropriate that this third series of *Papers in Applied Linguistics* be dedicated to chronicling *Advances in Language Testing*. The series will first survey the state of the art and then present theoretical, practical, and technical articles that record its progress. Each issue will have a specific theme; the series as a whole is meant to provide a means for continuing communication among all those concerned with language testing, whether as users, practitioners, or theorists.

The first three fascicles have a special history which should be mentioned here. They include original and revised versions of articles commissioned to appear in a volume intended to be called *Current Trends in Language Testing*. The original publisher's difficulties left the manuscripts in limbo for some time, and the size of the enterprise discouraged other publishers from taking up the project. Written as a survey of the state of the art, they are a good start for the present series.

Bernard Spolsky
April 1978

Introduction

This fascicle is intended to give some general notion of the state of the art in language testing by describing and reviewing several recent major tests. Its intention is to be representative rather than exhaustive: there are many other important tests that might have been included and that can be reviewed in future volumes in this series. Some brief words of introduction might explain the specific choices.

The IEA International Studies of Achievement have been major undertakings in the area of comparative education, seeking to show cross-nationally the varying factors that account for school success. When Professors Carroll and Lewis were asked to coordinate their respective sections of the Study, in the areas of French and English as foreign languages, they were faced with challenging problems in designing valid and practical tests. The account of their efforts and successes is a good description of the state of testing technology.

These two papers are written by the test developers. The others are much more in the nature of reviews, although sometimes reviews by insiders. The Test of English as a Foreign Language (TOEFL) was developed to meet the need for a standardized test for the many thousands of foreign students who apply each year to American universities. With changing administrative structure, the test has become a basic service for universities and a model for many others. Both Professor Oller and I served at different times on the Committee of Examiners, a role that permitted us to learn and comment, and this experience helped us to write our review. Dr. Paul Angelis, who is currently director of research for TOEFL, consented to bring our account up to date.

One of the most influential but least described language tests is the oral interview of the Foreign Service Institute. A brief description by Frank Rice appeared in the *Linguistic Reporter* in 1959 (Vol. 1, No. 2, p. 4); a longer one, by Claudia Wilds, appears in *Testing Language Proficiency* (Jones and Spolsky, 1975). Professor Jones now gives a detailed review of the test and its significance.

Each of the papers mentioned so far (except for Paul Angelis' update) were originally commissioned for *Current Trends in Language Testing* and revised for this publication. I have thought it appropriate to add one more review that illustrates some new developments. The Bilingual Syntax Measure, developed by Marina Burt and Heidi Dulay, and reviewed here by Professor Ellen Rosansky, has several interesting features: (1) it is based on a later linguistic theory than the other tests, (2) it has come to play a much debated role in a great deal of research in second language acquisition, and (3) it is being used by many bilingual education programs.

From these descriptions and reviews, one can see the developing field of language testing, with its clear double debt to the language sciences and to psychometrics, and its clear relevance to education.

Bernard Spolsky
April 1978

Contents

Preface . . . . *iii*

Introduction . . . . *v*

Language Proficiency Tests Developed for the IEA International Study of Achievement in French as a Foreign Language *1*
John B. Carroll

The IEA International Study of Achievement Test of Proficiency in English as a Foreign Language *49*
E. Glyn Lewis

The Test of English as a Foreign Language (TOEFL) *92*
John W. Oller, Jr. and Bernard Spolsky

TOEFL in Recent Years *101*
Paul J. Angelis

The Oral Interview of the Foreign Service Institute *104*
Randall L. Jones

A Review of the Bilingual Syntax Measure *116*
Ellen J. Rosansky

Bibliography *136*

Language Proficiency Tests Developed for the IEA International Study of Achievement in French as a Foreign Language*

John B. Carroll, Educational Testing Service

One of the subject matters selected by the International Association for the Evaluation of Educational Achievement (IEA)[1] for its comparative studies was French as a Foreign Language, along with English as a Foreign Language (treated in another chapter in this volume). The decision of the IEA Council to investigate achievement in a language other than English arose from the consideration that the English language occupies a somewhat unique position as a world language. In nearly all the non-English-speaking countries participating in the IEA project, it is either the principal second language or the most important foreign language. Motivation for the study of English in most of these countries possibly has a special character. It was therefore felt that valuable insights into foreign language teaching, whatever the language, could be gained by studying achievement in some other language than English that would be more representative of foreign languages that are offered in school curricula primarily for purposes of general education.

French was the obvious choice. It is generally an important foreign language, if not the most important one, in school curricula of the English-speaking IEA countries, and the most important second foreign language (after English) in those of the non-English-speaking countries. It was thought, further, that comparisons between results obtained in the English as a Foreign Language study with those obtained in the French as a Foreign Language study might reveal significant differences in teaching approaches and in types and levels of achievement that could be associated with the respective positions of the two languages in the various countries, and possibly also with the linguistic differences between English or French, on the one hand, and the mother tongues of the respective countries, on the other.

An International Committee on French as a Foreign Language was appointed by the IEA Council and began its work as early as November 1966. Besides the writer, who has served as its chairman, the members of the committee have been Clare Burstall (England), Guy Capelle (France and USA), John L. D. Clark (USA), Wallace E. Lambert (Canada), and H. N. Simai (Iran). This committee developed the overall design of the French as a Foreign Language study and supervised the construction of tests and other instruments and the eventual analysis and reporting of the results. National committees on French as a Foreign Language were formed in some of the participating countries to provide advice and information to the International Committee. The actual administration of the study (including selection of samples, collection and processing of data, and much of the statistical analysis) was performed by the IEA organization with the help of research groups ("national centers") in the respective countries. During the period February 1971 to January 1972, tests and other instru-

---

*Revised and updated by the author and reprinted by permission of the International Association for the Evaluation of Educational Achievement from John B. Carroll, *The Teaching of French as a Foreign Language in Eight Countries*. Stockholm, Almqvist and Wiksell and New York, John Wiley & Sons, 1975, pp. 47-111.

ments were administered to nearly 30,000 students in eight countries: Chile, England (including Wales), the Netherlands, New Zealand, Scotland, Sweden, Rumania, and the United States.

The purpose of the IEA studies has been not simply to compare achievements of different countries in different subject matters but rather to draw inferences about the ways in which factors in the individual student, in the school, and in the community at large promote or retard educational achievement. From such inferences, if they could be validly made, it has been hoped that important implications for educational policy and practice could be derived. The detailed statistical and other analyses that have been made of the French as a Foreign Language data for these purposes will not be presented here even in summary form, since the focus of the present volume (and thus of the present chapter) is on problems of foreign language *testing*.[2]

TEST DESIGN CONSIDERATIONS

In the IEA French as a Foreign Language project, tests and questionnaires of various sorts were the chief means of data collection. While it was not a primary purpose of the study to develop tests, it was necessary to develop tests that would have the highest possible reliability, validity, and practicality within the limitations of the project's resources and within the constraints imposed by the international character of the study. Test scores were to be used as criteria for assessing achievement in French and thus for drawing conclusions about factors contributing to achievement. The validity of these conclusions would depend, in large part, on the adequacy of the tests in assessing proficiency and achievement.

Objectives of the Teaching of French

From the start of the project, a prime consideration was that the tests should assess achievement in those aspects of competence in French that were regarded as most valuable, and most emphasized, in the various participating countries. These countries were asked to supply information about their "national objectives" in teaching French, with indications concerning the relative amount of emphasis in teaching the four skills (reading, writing, speaking, and listening) and the kinds of vocabulary and grammar items the students were expected to have mastered at various stages of instruction. From what information was supplied (some national centers were puzzled about our request, saying in effect that "French is French" and that there does not need to be differentiation of objectives), the clear impression was gained that the objectives of the teaching of French are highly similar across countries. In the main, French is taught for its cultural value, but it is also recognized as having practical values for many students. Because students vary widely in their motivations, an "omnibus" approach is adopted, with the hope that the student will attain as much proficiency as he can in both the spoken and written aspects of the language, and in both receptive and productive skills. Considerable emphasis is placed on the appreciation of French culture and the reading of French literature, particularly at advanced levels of instruction.

In view of these considerations about the objectives of the teaching of French, the International Committee decided that a more or less standard, but broad-ranging, battery of proficiency tests of French reading, listening, speaking, and writing could be prepared that would be about equally appropriate for all the countries that agreed to participate in the study. The content of the tests would feature the high-frequency elements of French vocabulary and grammatical structure commonly taught in school courses, avoiding specialized vocabularies of particular disciplines, subject matters, or occupations. While some members of the Committee hoped that it would be possible to test for the attainment of such cultural objectives as sensitivity to language structure, or knowledge of French history, art, and customs, it was felt that tests for these objectives might be difficult to construct and would, in any case, lengthen the test batteries beyond what would be feasible to administer. In the main, then, the tests were to be proficiency tests of competence in the French *language*.

As proficiency tests, the instruments were to be mainly of the "global" type, as opposed to the "discrete-point" type in which particular elements of the French language were to be the target of testing (Carroll, 1961, 1968). That is, the student's response to an item was usually to be a complex function of his total degree of competence in French, not his knowledge of any particular item of vocabulary, morphology, or syntax. This principle was generally followed in the Reading and Listening tests and in the "fluency" portions of the Speaking and Writing tests. In the pronunciation portion of the Speaking Skill test, however, and in one portion of the Writing test, it could be said that the "discrete-point" principle was followed, in that the student response was dependent usually on his competence with a particular sound or grammatical form of French. Nevertheless, the items tested were regarded as essentially a sample of a much larger stock of potential items, and these tests were scored for the total number of correct or acceptable responses rather than for any particular pattern of item responses. As will be seen, scores obtained in this way often correlate highly with scores from tests of a more "global" character.

Furthermore, there was little attempt to differentiate between competences in separate aspects of the language, such as lexicon and syntax. A frequent criticism of many foreign language tests of passive skills (reading and listening) is that successful performance on them depends too heavily on vocabulary knowledge and not enough on knowledge of morphology and syntax. This criticism may possibly be leveled at the IEA Reading and Listening tests, although there was a conscious effort to avoid making the tests depend too much on vocabulary. Knowledge of morphology and syntax comes to the fore, in any case, in some of the writing and speaking tests, for it is difficult for the student to *produce* acceptable sentences unless he has the requisite mastery of French grammar.

Target Populations

An important aspect of the design of every IEA study has been the designation of several distinct populations, at specified age or educational levels, which would be sampled in order to draw conclusions about factors accounting for differences in achievement among students, among schools, or among countries. An attempt was made to make the definitions of populations as uniform as possible. The study of French as a Foreign Language followed the pattern of other IEA studies, except that since French is usually an optional subject, the populations were restricted to students who had actually studied French or who were studying it at the time of testing. The specific definitions of populations adopted for the study were as follows:

Population I:	All students aged 10 in full-time schooling currently studying French.
Population II:	All students aged 14 in full-time schooling currently studying French.
Population IV:	All students in the pre-university year grade in full-time schooling who are currently studying French and have studied it for at least two years before the present academic year.
Population IVS:	The subpopulation of Population IV consisting of students who are "specializing" in French.

The reader should not be concerned with the fact that no Population III was defined; Population III was a "national option" that could be variously defined by national centers for their own purposes, if they so desired. Population IVS was also a "national option," but it entailed a special series of advanced-level tests. Tests for this population were, in fact, prepared and pretested, but, as matters turned out, no country elected to give the tests in the final field testing; thus, no data can be reported for these tests.

The general rationale for these definitions of populations has been described in other IEA publications (Walker, 1976). Population I was of particular interest in the French as a Foreign Language study because of concern with the problem of whether foreign languages can and should be taught at an early age. Populations II and IV span the range of students usually exposed to foreign language instruction at the secondary school level. Population IVS was defined in such a way as to assess how much achievement could be expected of students specializing in French at the secondary school level.

From the standpoint of test design, these definitions of target populations had implications for the levels of proficiency that the tests should be built to measure. It was felt that in order to provide sufficiently reliable and valid measures, separate batteries of tests would be required for each level, even though there might be some overlap among them. In fact, overlap of testing materials would be positively desirable in order to obtain a scaling of the scores that would be as meaningful as possible across populations and thus to access progress from level to level. At the same time it was thought desirable to provide a fairly wide range of test item difficulty at each level because students sampled from a given target population might actually represent a rather wide range of proficiency. In order to plan the levels of test difficulty, certain assumptions were made concerning the average level of proficiency for the typical student in each population. These assumptions were made on the basis of information from participating countries as to the grade levels when French instruction usually began, and the likelihood that a student might remain in French instruction over many grade levels.

In the case of Population I (10-year-olds), it was assumed that the typical student would have had only a year or two of "minimal" exposure (say, 15-30 minutes per day) to simple French vocabulary and grammatical structure.

For Population II (14-year-olds), it was assumed that students would have had two, or in some cases three, years of standard French secondary school courses.

Students at Population IV (pre-university year students) were expected to have had three or four years of secondary school instruction in French, and those defined in Population IVS (specialists) were expected to have had even longer and/or more intensive training in the language.

Practical Considerations

In previous IEA studies, it had been learned that the tests must be as short, simple, and uniform as possible, from the standpoint of test administration, answer-rendering, and scoring, if the cooperation of school administrators and teachers was to be obtained, and if the enormous quantities of data generated were to have any meaning. The testing of French as a Foreign Language (like that of English as a Foreign Language) presented special difficulties. While tests of reading skill could be patterned on traditional multiple-choice paper-and-pencil test formats, tests of listening skill might very well involve the use of tape recordings (unless it were decided to rely on the teachers themselves reading the French scripts for the test administration, and this was felt to be unwise). Use of tape-recorded student responses in a speaking test would involve not only the recording equipment but also much handling and labeling of tapes, and to obtain comparable, reliably scored data from different countries, central scoring of these tapes was almost imperative. Certain varieties of writing tests would also involve central scoring of responses. Despite the anticipated difficulties, the International Committee decided to specify the use of tape recordings in the Listening and Speaking tests, and to prescribe central scoring of student responses in Speaking and Writing tests wherever subjective judgments had to be made.

It was also desirable to keep total testing time as short as possible, while still maintaining satisfactory test reliabilities and minimizing the effect of any extraneous speededness in the tests. Time limits and pacing of items were to be set in such a way that the student would have a reasonable amount of time to deliberate about his answer and so that (in tests where the student paces his own work) most examinees could attempt every item. The extensive pilot testing of the testing materials made it possible to make reasonable decisions about all these matters when the final versions of the tests were put together. The pilot test materials included substantially more items than were needed in the final forms of the test, so that items of low validity could be either revised or discarded completely. Also, it was possible to calibrate testing times on the basis of observations regarding students' rate of work in the pilot testing.

Of some relevance to the design of the tests was the fact that they were to be administered internationally, to students of a variety of mother tongues. Students of contrastive linguistics might regard this situation as a golden opportunity to test hypotheses concerning the effect of linguistic differences on test-item performance. However, no special effort was made to develop or select test materials for such purposes--not only for lack of time, but also on account of the lack of relevant contrastive analyses. In any case, at the outset of the study, there was no accurate information as to which countries would agree to participate in the study; any special effort that might have been put into designing items to test hypotheses concerning particular pairs of languages (French-Persian, French-Swedish, etc.) might have been wasted. On the other hand, it may well be that the data may lend themselves to study from the standpoint of contrastive linguistics (see below, pp. 37-39).

Practical considerations dictated that wherever it seemed reasonable to do so, the tests should employ "objective" techniques whereby the student's response would be merely the selection from among a number of presented alternatives (as in the standard "multiple-choice" test, or as in a "true-false" test where there are essentially two alternatives). For most of the tests, students were to indicate their answers by marking appropriate positions on an "MRC" card. (The MRC card is the same size as a standard IBM card, and a completed card can be scored by machine.)

It was realized that objective testing techniques were not widely used in some countries that might participate in the IEA French study, i.e., that they might be somewhat unfamiliar to many prospective examinees. Thus, variations in "test-wiseness" might possibly cloud the overall results of the study. Nevertheless, experience in previous IEA studies, in subject matters other than French, seemed to indicate that even students unfamiliar with objective testing procedures (multiple-choice items, separate answer cards, and the rest) could adapt themselves to them readily. In the present study, the correlations between scores on "objective-type" tests and scores on free-response, subjectively-scored tests are often so high that it is difficult to believe that testing procedure ("objective" vs. "subjective") made any practical difference in the results.

THE CONSTRUCTION OF THE FRENCH PROFICIENCY TESTS

At the outset of test construction activities, several hundred provisional items were assembled; many of these came from existing test files in England (at the National Foundation for Educational Research in England and Wales) or in the United States (at Educational Testing Service), and others were newly constructed. Items drawn from existing test item files were "nonsecure," i.e., they had for one reason or another been released from the files of items that are maintained for use in "secure" examinations not available for public inspection. Many of these "nonsecure" items had been repeatedly used previously, and extensive item analysis data (on item difficulty and discrimination power) were available. Even so, subsequent use of these items in pilot testing under IEA auspices, and scrutiny of the items by members of the International Committee or by French experts in participating countries, often disclosed ways in which the items needed improvement. Sometimes, for example, the phraseology of the items was found to be excessively academic or actually incorrect from the standpoint of modern, idiomatic French as spoken in metropolitan France. Sometimes, also, it was found that while a given multiple-choice item had effective "distractors" (incorrect alternatives) in the United States, these distractors did not work equally well in all countries, possibly because of different instructional materials, or possibly because of different kinds of contrasts between French and the mother tongue that existed in the non-English-speaking participating countries. In all test construction activities, an effort was made to select and develop items that would be about equally effective, as judged by item validity statistics (using total test score as criterion) in the various participating countries.

Also, an effort was made to balance the items with respect to the phonological, lexical, grammatical, and semantic features of French that were to be tested. For example, it was found that most of the stimulus sentences in an early form of the Listening Comprehension test were affirmative statements, with negative and interrogative structures underrepresented. These deficiencies were remedied by retrieving replacement items from item files, by making appropriate modifications in the existing items, or by creating and trying out entirely new items.

In general, items in multiple-choice tests (Reading and Listening) were placed in rough order of increasing difficulty.

As indicated above, the final forms of the tests evolved through a process of pilot testing, item analysis, and revision through at least two cycles. It would be tedious to relate all the details of this process; I will content myself with a description of the final forms of the tests.

Table 1 presents an overview of the final plan of the proficiency tests as devised for the several populations. In this table, items are identified by types; each type of item will be illustrated and discussed below. Where tests are identical over populations (either in whole or part), this is so indicated in the table. Selected results from the final testing will be presented to illustrate item characteristics.

Listening Tests

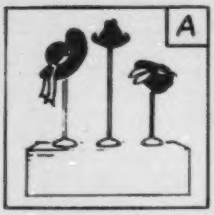
Listening tests were constructed at four different levels of difficulty so as to be appropriate for Populations I, II, IV, and IVS, respectively. However, some items, called "anchor items," were identical from one form to another so that score scales could be vertically equated. The tests were designed to be administered with tape recordings of the stimulus materials. The portions in French were recorded in Paris by Guy Capelle (a member of the International Committee) and a group of other native speakers that he assembled.[3] Some of the items involved dialogues with two or more speakers, and some parts of the tests used male and female speakers for alternate items. Copies of the master tape were sent out to participating countries with indications as to where instructions translated into the respective mother tongues should be inserted in making submaster tapes for these countries. The item numbers, however, were to remain in French.

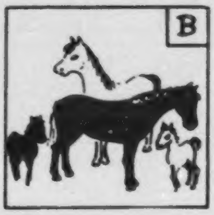
In every case, the student had a printed test booklet containing further materials on which he was to base his answer--either sets of pictures, or sets of alternatives printed in French. Pictorial materials were used where it was considered desirable to avoid measuring the student's ability to read printed French. The fact that many of the tests did, nevertheless, use printed French alternatives means that to some extent performance on the listening tests may have depended on reading comprehension skill. It was assumed that students at the more advanced levels had enough skill in reading comprehension to minimize this factor.

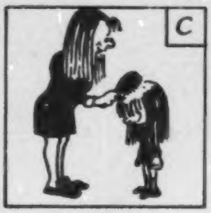
Pictorial Items. The Listening test for Population I was composed exclusively of this type of item, and a few such items were employed in the Population II and IV tests.

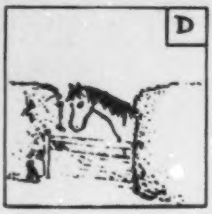
An item (No. 5) from the Population I Listening test is as follows:

Tape: *Ce sont des chevaux.*
(Time: 1 sec., followed by silence of 18 sec.) (woman)









Obviously the discrimination to be made is between *chevaux* ("horses"), *cheveux* ("hair," plural), and possibly *chapeaux* ("hats"). The percentages of USA

Table 1

Overview of the Final IEA French Proficiency Tests

Skill, Item Type	Population I No. of Items	Population I Max. Score	Population I Admin. Time (min.)	Population II No. of Items	Population II Max. Score	Population II Admin. Time (min.)	Population IV No. of Items	Population IV Max. Score	Population IV Admin. Time (min.)	Population IVS No. of Items	Population IVS Max. Score	Population IVS Admin. Time (min.)
Listening												
Pictorial	35	35		11	11		7	7		--	--	
Sentences	--	--		12	12		9	9		12	12	
Conversations	--	--		9	9		8	8		8	8	
Passages	--	--		8	8		5	5		8	8	
Long Conversations	--	--		--	--		11	11		9	9	
Total	35	35	20*	40	40	25*	40	40	25*	37	37	25*
Reading												
Pictorial	35	35		--	--		--	--		--	--	
Vocabulary	--	--		18	18		17	17		15	15	
Texts	--	--		17	17		22	22		21	21	
Total	35	35	20	35	35	25	39	39	25	36	36	25
Speaking												
Pronunciation	16	29	....	Same as Pop. I		....	Same as Pop. I			--	--	
Structural Control	10	40		10	40	....	Same as Pop. II			--	--	
Oral Reading	--	--		1 passage	24							
Fluency	1	**		2	**	....	Same as Pop. II			--	--	
Total		**	15*		**	25*		**	25*			--
Writing												
I Grammar	--	--		25	25	15	26	26	(15)†	--	--	
I Verbs, Modifiers	--	--		7	20		6	16		--	--	
					45			42				
II Directed Composition	--	--		1	**	10	1	**	(25)†	--	--	
Total			--		**	25		**	40			--

*Times for these tape-recorded or self-paced tests are approximate. The administration times given do not include time for giving instructions.

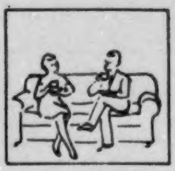
**See text for details concerning the scoring of the Speaking and Writing Fluency tests.

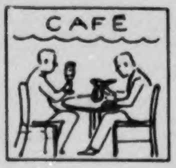
†The student could proceed to the Directed Composition, if he wished, before the time limit for the first two sections (Grammar, Verbs & Modifiers) was up.

Population I students (*N*=206) responding to the respective alternatives were 15, 45 (correct), 22, and 18; the item validity (point-biserial correlation with total score excluding this item) was .575. (Population I tests were given only in the United States.) This was therefore a highly valid item, and the distractors were about equally effective.

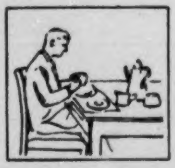
The style of drawing for the pictorial items in the Population II and IV tests was somewhat more schematic. An item that was used as an "anchor item" in both test levels was as follows:

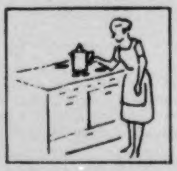
Tape: *Après le dîner, les parents prennent du café.*
(Time: 3 sec., followed by silence of 12 sec.)


(A)



(B)


(C)

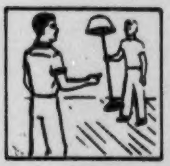

(D)

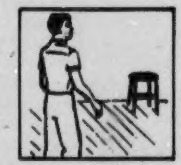
Item analysis results for this and other items illustrated here, drawn from the data of the final testing, are given in Table 2.

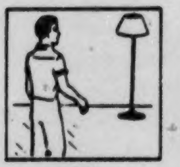
The above item (Item 3, Population II test, identical to item 5, Population IV test) was relatively difficult for Population II but relatively easy for Population IV, and generally it was a highly discriminating item. Poorer students had a strong tendency to select alternative B.

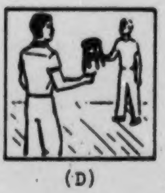
A fair amount of subtlety can be incorporated in a pictorial item. For example, here is an item (No. 6, Listening Population IV test) that was difficult even for Population IV:

Tape: *Le fils s'est trompé. Le père avait dit: "Apporte-moi la lampe."*
(Time: 4 sec., followed by silence of 10 sec.)


(A)


(B)


(C)


(D)

From the item analysis results (Table 2) we may see that, like the boy in the picture, most of the examinees were deceived, thinking (A) was the correct alternative.

Sentences. In this type of item, the student was told that he would hear a remark or question in French, and that he was to select, from among the four choices in the test booklet, the response that would most likely be made to the remark or question. An item appearing on both test levels (No. 17, Population II, and No. 9, Population IV) was as follows:

Tape (woman's voice): *Qu'est-ce que vous donnez à manger à votre chien?*
(Time: 1½ sec., followed by silence of 10 sec.)

Booklet: (A) *Oui, souvent.*
(B) *Tres gentil.*

Table 2

Selected Item Analysis Data for Listening Tests, by Country and Population (Percentages Marking Options or Omitting the Item, and Point-Biserial Correlations with Total Score Based on the Remaining Items)[1]

	Population II						Population IV					
Country	A	B	C	D	Omit	r_{pb}	A	B	C	D	Omit	r_{pb}
			Item 3						Item 5			
Chile	--	--	--	--	--	--	22.7*	15.0	21.9	30.6	7.7	.435
Eng.	40.3*	38.3	12.2	8.6	0.7	.379	87.5*	1.3	5.5	4.7	0.0	.270
N.Z.	41.6*	37.5	11.2	9.5	0.2	.374	78.3*	7.3	9.4	4.7	0.2	.162
Rum.	41.9*	15.4	32.5	0.9	4.2	.090	64.0*	19.2	2.6	1.3	3.8	.452
Scot.	46.0*	35.0	9.9	9.0	0.0	.498	79.5*	7.5	9.0	4.0	0.0	.240
Swe.	--	--	--	--	--	--	80.8*	6.4	9.8	2.9	0.2	.322
USA	37.7*	48.3	8.7	4.6	0.6	.423	75.7*	11.2	9.3	3.5	0.1	.380
			[Item not in Pop. II Test]						Item 6			
Chile							57.5	5.0	15.7	5.5*	14.2	.043
Eng.							45.4	0.0	1.5	51.7*	0.4	.392
N.Z.							55.7	8.2	10.4	23.8*	1.8	.088
Rum.							52.5	1.3	0.0	32.0*	2.6	.302
Scot.							67.8	1.6	2.8	27.5*	0.3	.287
Swe.							64.4	2.1	4.6	28.4*	0.6	.418
USA							69.5	2.1	7.0	20.2*	1.0	.391
			Item 17						Item 9			
Chile	--	--	--	--	--	--	8.1	10.5	36.1	21.1*	22.6	.459
Eng.	11.4	13.0	37.8	37.3*	0.2	.534	0.5	2.3	6.1	90.1*	0.0	.403
N.Z.	16.9	23.4	43.0	16.0*	0.7	.113	10.4	5.4	34.5	49.0*	0.8	.384
Rum.	2.5	1.7	0.0	81.3*	2.6	.813	0.0	0.0	0.0	87.0*	1.3	.866
Scot.	12.0	13.3	43.8	29.9*	1.0	.409	3.8	6.4	20.0	69.4*	0.5	.484
Swe.	--	--	--	--	--	--	4.2	7.1	21.1	66.7*	0.9	.499
USA	10.7	20.6	41.7	25.5*	1.1	.354	3.9	8.9	21.2	65.2*	0.7	.589
			Item 25						Item 19			
Chile	--	--	--	--	--	--	7.4	39.4*	21.0	22.6	6.5	.435
Eng.	7.7	51.2*	21.3	18.5	0.9	.418	0.7	93.7*	1.0	3.5	0.0	.283
N.Z.	9.1	41.5*	28.6	20.7	0.1	.217	1.0	76.7*	9.4	12.7	0.3	.147
Rum.	6.8	75.3*	0.0	2.6	2.6	.830	0.0	84.5*	0.0	0.0	2.6	.951
Scot.	9.8	42.8*	26.3	20.5	0.6	.237	1.2	76.2*	12.5	9.5	0.6	.165
Swe.	--	--	--	--	--	--	1.0	74.2*	10.8	13.1	0.7	.272
USA	10.6	38.4*	30.5	19.3	0.9	.236	2.6	72.5*	11.2	13.0	0.5	.336

Table 2 *(continued)*

	Population II						Population IV					
	A	B	C	D	Omit	r_{pb}	A	B	C	D	Omit	r_{pb}
			Item 31						Item 29			
Chile	--	--	--	--	--	--	11.0*	32.9	23.5	5.5	22.9	.118
Eng.	14.5*	53.3	15.4	13.6	0.3	.225	87.4*	4.8	4.6	2.2	0.0	.275
N.Z.	7.6*	54.1	25.2	13.0	0.1	.020	37.0*	38.5	13.3	9.5	1.6	.464
Rum.	53.9*	17.9	5.1	8.6	0.8	.558	81.9*	1.3	0.0	0.0	2.6	.914
Scot.	9.6*	51.8	23.2	15.0	0.4	.082	52.7*	27.2	12.2	7.3	0.6	.462
Swe.	--	--	--	--	--	--	32.6*	30.3	23.1	12.3	1.6	.501
USA	8.1*	50.3	23.5	17.0	0.7	.118	30.4*	35.7	16.8	15.6	1.1	.588
			Item 35						Item 30			
Chile	--	--	--	--	--	--	23.2	13.0*	20.3	33.4	5.2	.192
Eng.	24.8	19.6*	17.5	33.1	1.5	.093	21.0	68.3*	1.8	7.9	0.0	.328
N.Z.	24.9	18.4*	15.5	39.7	1.5	.014	24.2	46.7*	9.1	19.2	0.7	.183
Rum.	2.6	69.3*	5.1	8.6	0.0	.561	9.0	66.5*	2.6	3.8	3.8	.641
Scot.	23.6	23.0*	16.5	36.0	1.0	.035	27.1	53.2*	5.4	14.1	0.2	.244
Swe.	--	--	--	--	--	--	14.9	64.8*	11.3	8.8	0.2	.388
USA	24.3	19.4*	14.7	40.0	0.7	.117	19.7	48.3*	13.4	18.1	0.3	.506
			Item 36						Item 31			
Chile	--	--	--	--	--	--	24.6	23.1	10.2*	10.7	24.5	.304
Eng.	24.1	16.6	33.2*	19.7	2.7	.199	19.1	7.8	49.7*	20.0	2.4	.342
N.Z.	29.3	16.0	26.0*	26.2	1.8	.015	24.7	11.0	28.4*	34.2	1.7	.229
Rum.	35.1	4.3	41.0*	5.1	0.0	.339	2.6	17.9	39.7*	20.5	5.1	.358
Scot.	24.9	15.0	31.6*	25.6	2.2	.108	21.9	12.0	31.9*	31.6	2.5	.156
Swe.	--	--	--	--	--	--	34.0	10.2	32.5*	21.5	1.7	.312
USA	31.4	14.7	23.9*	27.9	1.1	.066	34.7	7.0	29.4*	27.5	1.0	.288

*Percentage marking the option keyed as correct.

[1] Data for items common to the Population II and IV tests are displayed alongside each other. If the percentages do not sum to 100, the remainder is the percentage not reaching the item. The *N*'s on which all data are based are as follows:

	Pop. II	Pop. IV
Chile	--	173
Eng.	455	180
N.Z.	1790	361
Rum.	117	77
Scot.	834	974
Swe.	--	1742
USA	1927	1370

Booklet: (C) *C'est un animal.*
(D) *De la viande.*

(C) was a particularly attractive alternative to many of the poorer students.

Short Conversations. The student was told that he would hear a short conversation between two people: that "one person will ask a question or make a statement, and the other will reply." After each conversation, he was to mark the statement (among four alternatives) "which is correct according to what has been said." An item that appears on both test levels (No. 25, Population II, and No. 19, Population IV) is as follows:

Tape: (man) *Voici la valise que vous avez laissée ici, mademoiselle.*
(woman) *Merci beaucoup, monsieur.*
(Actual time: 5 sec., followed by silence of 13 sec.)

Booklet: (A) *La femme n'a pas de valise.*
(B) *L'homme lui donne sa valise.*
(C) *La femme achète une valise.*
(D) *L'homme cherche la valise.*

This was of appropriate difficulty for Population II, yet also valid for Population IV. The distractors were generally quite effective.

Passages. The student was told that he would hear a short broadcast announcement, and be asked a question about what was said, selecting the "best answer" from among the four choices printed in his booklet. An item appearing on the tests at both levels (No. 31, Population II; No. 29, Population IV) is as follows:

Tape: (man) *Un ouragan s'est abattu non loin de Londres et a endommagé tout un village. De gros arbres ont été déracinés par le vent, et les toitures de plusieurs bâtiments ont été arrachées par des tourbillons violents.*
(Actual time: 14 sec.)

Tape and booklet: *De quoi s'agit-il?* (followed by silence of 12 sec.)

Booklet: (A) *D'une tempête.*
(B) *De la modernisation d'un village.*
(C) *D'un accident d'avion.*
(D) *D'une bataille.*

For some samples in Population II, this was of such difficulty as to be almost invalid, but it was quite valid for samples in both populations for which it was of appropriate difficulty.

Long Conversations. The student was told that he would hear "rather long conversations or dramatic scenes," and after each one would be asked several questions. One of the conversations, and two items based on it, that appears in tests at both levels (Items 35 and 36, Population II; Items 30 and 31, Population IV) is as follows:

Tape: (man) *Ah! Marie, tu rentres d'une promenade?*
(girl) *Oui. Il fait très beau ici à la ferme ce matin.*
(man) *Qui t'a donné ces belles pommes?*
(girl) *Je les ai cueillies moi-même en sortant du petit bois.*
(man) *J'oubliais de te dire qu'il y a une lettre pour toi dans ta chambre.*
(girl) *D'où vient-elle?*
(man) *De Paris.*
(girl) *Mes parents me demandent de revenir à la maison. J'en suis sûre.*
(man) *J'espère que tu ne vas pas partir tout de suite. Nous serions désolés.*

(girl) *Moi aussi. En tout cas, je passerai toute la journée ici avec toi et tante Louise. Je te le promets.*
(man) *Ça nous fera bien plaisir.*
(Actual time: 35 sec.)
(another man) *Voici les questions.*
Où se trouvent ces deux personnes? (followed by pause of 12 sec.)

Booklet:
(A) *Dans un hôtel.*
(B) *A la campagne.*
(C) *Au marché.*
(D) *A Paris.*

Tape: *Qu'est-ce que Marie a rapporté de sa promenade?* (pause, 12 sec.)

Booklet:
(A) *Une lettre.*
(B) *Des branches.*
(C) *Des fruits.*
(D) *Des fleurs.*

The reader may refer to Table 2 to see how these items functioned.

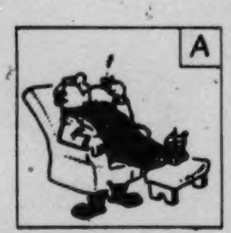
Scoring of the Listening Tests. Two types of scores were derived from the results: a total raw score (the number of items correctly answered), and a formula-corrected score (the number of correct answers minus one-third the number wrong, not counting "omits"). The formula-corrected score was intended to correct for differential tendencies to omit items or to guess answers. Students were encouraged to "answer every question even though you may not be sure that your answer is correct." Very small percentages of students omitted answers.

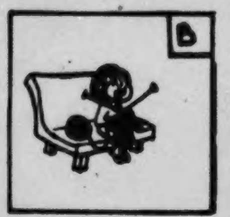
In composing the tests, it was intended that there be a considerable range of item difficulty in the light of the expected levels of achievement in the different populations. An analysis of the item difficulty statistics resulting from the final testing is given later in this chapter.

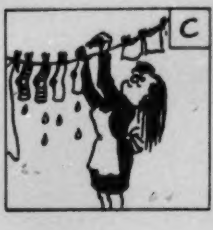
Reading Tests

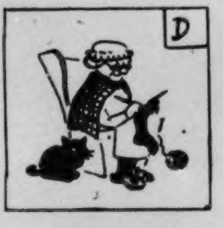
Reading tests were composed of various types of items, described here. As in the case of the Listening tests, it was intended that there be a wide range of item difficulty in the test for each level.

Pictorial Items. Pictorial items were used only in the Population I Reading Comprehension test, which was composed exclusively of this type of item in order to make the student's answer depend solely on his comprehension of the "lead" and not on his comprehension of the meanings of the presented alternatives (assuming that he could readily interpret the line drawings). One of the more difficult items (No. 7) --an item with one of the highest item validities (.579) in the sample of 206 students in Population I tested in the United States--is as follows:









Elle tricote une chaussette.

No student omitted the item; the percentages answering the several alternatives were: (A) 5.8; (B) 18.9; (C) 21.8; (D, correct) 53.4. Most of the wrong answers, therefore, resulted from failure to interpret either *tricote* or *chaussette*.

Vocabulary Items. Vocabulary items occurred in the Reading tests for Populations II, IV, and IVS. They are here called "vocabulary" items because successful answering usually depended largely on knowledge of the meanings of one or two key words occurring (generally) in the item alternatives. The "lead" statements, however, gave considerable context and had to be comprehended in order to select the correct answer. Thus, a variety of lexical and syntactic knowledges were tested; exactly how an item would function for a particular student could not be specified in advance. An item for which Population II (No. 6) and IV (No. 2) results are given in Table 3 is as follows:

Jean revient à la maison. Sa mère lui dit: "Comme tu es méchant!
Tu as . . . ton pantalon!"

(A) *trouvé.*
(B) *acheté.*
(C) *porté.*
(D) *déchiré.*

On the assumption that the student comprehends the lead, his choice of the correct answer (*déchiré*, "torn") depends either on his knowledge of the meaning of that word or his ability to eliminate the remaining alternatives as *not* making sense in the context given. It is possible, however, that many students failed to understand the lead statement correctly, not knowing words such as *méchant* (naughty) and *pantalon* (trousers).

Texts. "Text items" consisted of printed passages of various lengths and difficulties about which the student was asked several questions (in French). An example, found on both Population II (Nos. 33-35) and IV (Nos. 28-30) tests, is as follows:

Lorsque Jacques Lafitte est venu à Paris, il s'est présenté chez M. Perregaux dans l'espoir d'obtenir une place.

Le banquier lui dit qu'il est dans l'impossibilité de satisfaire à sa demande, parce qu'il n'a besoin de personne.

Lafitte, découragé par ce refus, s'éloigne tristement, lorsqu'en traversant la cour de l'hôtel, il aperçoit une épingle par terre et il la ramasse. M. Perregaux, qui regardait par la fenêtre, le voit et en est frappé: le jeune homme, pense-t-il, doit avoir un esprit d'ordre et d'économie. Il le fait rappeler et lui dit qu'il peut compter sur une place dans sa banque.

33. *Pourquoi Jacques Lafitte s'est-il adressé à M. Perregaux?*
(28.)
(A) *Il cherchait un emploi.*
(B) *Il voulait quitter Paris.*
(C) *Il avait de l'argent à placer.*
(D) *Il désirait lui offrir une situation.*

34. *Qu'est-ce que M. Perregaux lui dit d'abord?*
(29.)
(A) *Qu'il ne pouvait pas lui donner de travail.*
(B) *Qu'il était satisfait de recevoir sa demande.*
(C) *Qu'il était possible de satisfaire tout le monde.*
(D) *Qu'il était sans argent.*

35. *Ce que le jeune homme a fait dans la cour a*
(30.)
(A) *fait rire le banquier.*
(B) *découragé le banquier.*
(C) *attiré l'attention du banquier.*
(D) *fait peur au banquier.*

Table 3

Selected Item Analysis Data for Reading Tests, by Country and Population (Percentages Marking Options or Omitting the Item, and Point-Biserial Correlations with Total Score Based on the Remaining Items)[1]

	Population II						Population IV					
Country	A	B	C	D	Omit	r_{pb}	A	B	C	D	Omit	r_{pb}
				Item 6						Item 2		
Chile	--	--	--	--	--	--	17.3	28.5	16.7	31.7*	5.7	.269
Eng.	16.7	18.4	26.9	37.1*	0.9	.537	0.1	0.2	0.3	99.0*	0.4	.271
Neth.	9.8	8.9	17.0	63.4*	1.0	.362	7.6	4.3	6.9	86.8*	0.3	.228
N.Z.	16.4	21.8	13.4	47.9*	0.5	.497	1.0	0.1	0.7	98.1*	0.0	.094
Rum.	6.2	3.7	7.9	81.6*	0.5	.541	0.0	0.2	0.1	99.7*	0.0	.118
Scot.	15.2	14.6	18.6	50.4*	1.2	.561	1.0	0.6	2.4	95.7*	0.3	.267
Swe.	--	--	--	--	--	--	8.0	7.6	11.5	72.2*	0.7	.485
USA	21.1	28.5	22.3	27.2*	0.8	.345	5.5	12.7	10.4	71.0*	0.4	.527
				Item 33						Item 28		
Chile	--	--	--	--	--	--	47.8*	9.3	21.7	13.0	3.7	.329
Eng.	25.2*	28.0	15.6	15.2	0.4	.426	96.5*	0.1	0.2	3.3	0.0	.319
Neth.	38.4*	18.1	21.6	8.4	1.2	.358	88.2*	1.0	0.7	9.5	0.1	.228
N.Z.	30.7*	17.3	18.7	22.5	0.7	.431	92.4*	0.1	0.1	7.3	0.0	.223
Rum.	57.8*	6.0	9.0	25.6	0.2	.473	74.4*	2.8	3.0	16.1	2.2	.541
Scot.	37.1*	18.1	14.3	19.3	0.0	.416	88.2*	0.3	1.2	10.2	0.1	.281
Swe.	--	--	--	--	--	--	87.0*	2.7	3.5	5.7	0.2	.333
USA	31.6*	22.5	21.0	18.4	0.6	.292	74.9*	3.1	7.8	12.9	0.2	.444
				Item 34						Item 29		
Chile	--	--	--	--	--	--	36.4*	27.2	16.7	8.4	5.6	.346
Eng.	26.1*	22.7	20.0	11.6	0.6	.324	97.8*	1.2	0.8	0.1	0.0	.328
Neth.	29.4*	26.8	17.4	10.4	0.9	.384	94.0*	3.8	0.8	0.4	0.1	.317
N.Z.	26.0*	29.1	22.1	9.5	0.6	.260	95.3*	2.9	1.1	0.6	0.0	.232
Rum.	80.6*	6.8	4.9	5.6	0.1	.535	85.1*	6.9	2.3	1.5	2.5	.519
Scot.	28.9*	25.5	20.8	9.9	0.8	.414	92.9*	3.5	2.8	0.7	0.1	.355
Swe.	--	--	--	--	--	--	81.7*	8.4	5.2	3.2	0.3	.499
USA	25.0*	29.4	25.4	12.5	0.5	.264	72.8*	13.5	7.3	4.5	0.7	.529
				Item 35						Item 30		
Chile	--	--	--	--	--	--	10.1	16.9	34.5*	22.8	8.1	.240
Eng.	12.2	15.9	27.5*	22.7	0.0	.322	0.0	0.7	98.1*	1.1	0.0	.304
Neth	10.7	23.4	30.4*	18.2	0.0	.289	0.8	5.9	89.0*	2.9	0.3	.339
N.Z.	11.4	20.6	35.2*	17.8	0.0	.331	0.3	2.1	96.4*	1.0	0.0	.102
Rum.	3.4	8.4	77.0*	8.8	0.0	.590	2.8	4.2	84.8*	3.6	2.5	.578
Scot.	9.8	18.6	32.5*	22.9	0.0	.354	0.9	6.6	89.1*	3.0	0.3	.344
Swe.	--	--	--	--	--	--	3.6	14.4	71.1*	8.5	0.8	.469
USA	15.3	24.6	26.1*	25.6	0.0	.218	4.2	10.9	74.1*	8.4	0.8	.510

*Percentage marking the option keyed as correct.

[1] Data for items common to the Population II and IV tests are displayed alongside each other. If the percentages do not sum to 100, the remainder is the percentage not reaching the item. The *N*'s on which all data are based are as follows:

	Pop. II	Pop. IV
Chile	--	1440
Eng.	2076	702
Neth.	1545	1753
N.Z.	1793	363
Rum.	2271	2247
Scot.	835	972
Swe.	--	1755
USA	4177	3069

In constructing these items, an effort was made to avoid problems that inexperienced test constructors often overlook, namely those of (1) making the questions in such a way that they can be successfully answered without reading the text on which they are based, and (2) making the items interdependent in such a way that an answer given for one item (whether correct or not) is itself a cue for selecting an answer in another item. Actually, the example given here is probably satisfactory in the first respect, but it is not wholly satisfactory in the second respect, in that answer (A) to the first item is related to answer (A) in the second item, and both these answers happen to be correct. Nevertheless, the item analysis data show the items to be generally satisfactory.

Time Limits. Pilot testing data indicated roughly how rapidly students would perform items of the types included in the Reading test. In general, the time limits were ample enough to permit most students to work through the test and attempt even the last item. (Table 4 gives the percentages of students in various samples who did not attempt the next-to-last item; statistics are given for the next-to-last item because similar statistics for the last item would include normal "omits" for that item.) Since the vocabulary and text sections were not separately timed or scored, the results give little indication concerning average reading comprehension speeds. The scores measure, rather, general ability to comprehend written French accurately when the student is given a reasonable amount of time in which to do so. The exact role of lexical as opposed to grammatical knowledge cannot be assessed, but it is probable that lexical knowledge was stressed, since it is difficult to identify items or passages in which knowledge of a particular grammatical feature of French would be critical in selecting a correct answer, whereas it is easy to find items or passages in which lexical knowledge would be critical.

Table 4

Percentages of Students Not Reaching the Next-to-Last Item

	Reading Population			Listening Population		
Country	I	II	IV	I	II	IV
Chile	--	--	28.6	--	--	21.1
England	--	19.1	3.5	--	3.7	1.0
Netherlands	--	15.1	12.7	--	--	--
New Zealand	--	12.7	9.9	--	1.0	1.9
Rumania	--	1.9	7.5	--	14.5	18.1
Scotland	--	14.2	9.4	--	1.6	0.7
Sweden	--	--	12.9	--	0.4	0.7
USA	1.0	7.1	11.0	0.5	1.5	1.6

Scoring. As in the case of the Listening tests, both raw and formula-corrected scores were derived from the results. The corrected scores were number correct minus one-third the number wrong, not counting "omits." The formula-corrected scores were intended to correct not only for differential guessing tendencies but also (at least partially) for the fact that some students did not reach the end of the test within the time limit. The instructions for the test encouraged guessing, however: the student was told "It will be to your advantage to answer every question even though you may not be sure that your answer is correct." These instructions are based on the assumption that the scores would be more accurate indications of competence if guessing tendencies were as nearly as possible equated across students, and on the assumption that a guess in the absence of *certain* knowledge is still likely to reflect *some* knowledge. (For example, in the vocabulary item illustrated above, we noted that the correct answer might be marked by a student on the basis of eliminating the wrong alternatives, even without being certain of the meaning of *déchirer*.)

Speaking Tests

In an effort to develop measures of competence in speaking French, tests of several somewhat different aspects of speaking competence were devised. The Pronunciation test and the Oral Reading test supposedly measured the ability to pronounce French with acceptable phonology; the Structural Control test sought to measure the student's ability to select appropriate vocabulary and syntax to answer questions about simple but controlled situations; the Fluency test was designed to obtain information about the quantity and quality of the student's response in French when he was asked to speak somewhat more at length in describing sequences of events or in pictured situations.

It should be mentioned that since the Speaking tests involved extra testing time and special arrangements for scoring, they were administered to relatively small samples of students in the various countries--actually, from 2.8 to 39.6 percent of the total samples tested. It was felt that the project would be subject to criticism if it did not obtain at least some information on actual spoken production competencies of students, since the Reading and Listening tests--generally given to all students tested, or to major portions of them--tested only receptive skills.

All Speaking tests were administered with the help of *two* tape recorders. One machine was used for playback of a copy of a master tape that was furnished to the tester; the other machine was used for recording student responses. Testers were required to follow directions regarding exactly how the tests were to be administered; this necessitated frequent manipulation of the start and stop controls of the machines, and careful attention to timing and to recording quality.

Generally, response tapes contained the responses of a number of students. These tapes were collected and sent to a central location (in Liège, Belgium), where they were scored and evaluated by a carefully-trained team of native speakers of French. Because of the volume of materials collected from the various countries (responses from a total of more than 2000 students in various populations), scoring activities took place over a number of months. The scoring team made efforts to maintain constant standards of evaluation across the various members of the team. Student tapes were normally scored by only one member of the team, but in order to obtain estimates of scorer reliability, some tapes were scored independently by two members of the team. Further information about scoring methods is given in connection with the discussion of each type of test. Information on the reliability of scoring is given later in the chapter (pp. 32-35).

Pronunciation Test. The same test was used for all populations, since the same phonological problems occur at all levels of instruction. It consisted of a series of short French sentences, each of which was heard rather carefully pronounced (by a native speaker of French) on a master tape *twice* in immediate succession, followed by a pause of about six seconds during which the student was simply to repeat what he had heard (recording the response on the student tape). There was no instruction to the student that he try to imitate the sounds exactly in his "best French pronunciation"; it was assumed that the instruction merely to *repeat* was sufficient to reveal whatever degree of accuracy the student was capable of. There were three practice items: *"Quelle heure est-il," "Elle est là-bas,"* and *"Il fait chaud,"* followed by 16 scored items that together contained a variety of French phonemes and problems of particular interest (e.g., certain vowels such as [y] and [œ]; nasalized vowels; liaisons; etc.). Scoring of the student responses was based on assignment of either 0 (not acceptable) or 1 (acceptable), to rendition of each of 29 particular phonemes or phonological features that occurred in the 16 sentences. Thus the scores could range from 0 to 29. It is impossible to indicate here exactly what standards of acceptability were used by the scoring team, but for the purposes of the study it was only necessary that these standards be reasonably consistent. It may be assumed that the standards were fairly rigorous since scoring was done by native speakers of

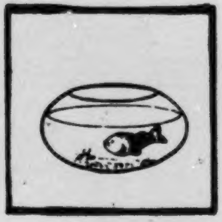
French. (Generally the scorer did not know the population to which a particular student belonged.)

Structural Control. This test was administered directly after the Pronunciation test. It was prepared in two forms, a relatively simple one for Population I, and a somewhat more difficult one for Populations II and IV. The test consisted of a series of pictures in the student's test booklet, accompanied by tape-recorded questions (in French) which were to be answered by the student in French (with the response recorded on the student tape). The student was told: "Listen carefully to each question, because it will only be asked once. Try to answer each question as soon as you have heard it. If you do not understand a question or cannot answer it, you can say *'Je ne sais pas.'* Try to answer each question with a whole sentence, not just a single word like *'oui'* or *'non.'*" There were two practice items, followed by ten questions spoken on the master tape just *once.* Each question referred to a corresponding picture in the test booklet. The test was strictly timed by the tape; a silence of about 15 seconds occurred on the tape after each question, allowing the student to respond.

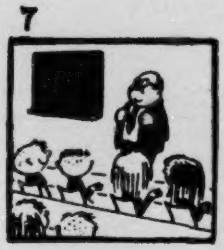
To give an idea of the types of items and levels of difficulty, here are Items 3, 7, and 11 from the Population I and II tests (the pictures are from the test booklet and the questions have been transcribed from the tape):

Population I

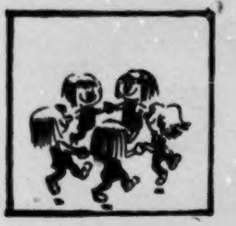
3



7



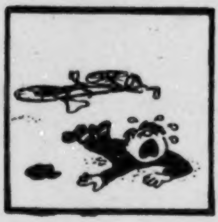
11



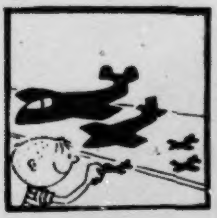
Tape: 3. *Qu'est-ce que c'est?*
7. *Où sont les enfants?*
11. *Combien d'enfants y a-t-il?*

Population II

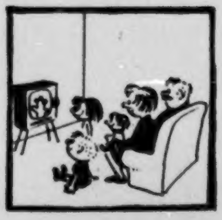
3



7



11



Questions: 3. *Qu'est-ce qui est arrivé au garçon?*
7. *Est-ce que il prend le grand avion?*
11. *Que font ces gens?*

It may be noticed that for the "yes/no" questions it was usually necessary for the student to deny the assertion and then indicate what was actually the case. Of course, since the student had to understand the question (always given in French), it could be said that this test was partly a test of listening comprehension, but the use of pictures may have minimized this problem.

Scoring of the test responses was based on a scale ranging from 0 to 4 for each response; thus the total score (for ten responses) could range from 0 to 40. The scale is as follows (here translated from the French):

0 - no response - "yes" or "no" - inappropriate or unintelligible response - "I don't know"
1 - an accumulation of serious errors *or* one serious error accompanied by one or several minor errors
2 - one serious error *or* an accumulation of minor errors
3 - one or two minor errors
4 - completely correct

Oral Reading. This test was given directly after the Structural Control test, but only to students in Populations II and IV. It consisted of a single paragraph (15 sentences, 165 words) of relatively simple French prose printed in the student's test booklet. The student was allowed a maximum of three minutes to look over the passage, after which he was told to read the passage "loudly and clearly, directly into the microphone." (Again, no special instructions were given as to how careful the pronunciation and intonation should be.)

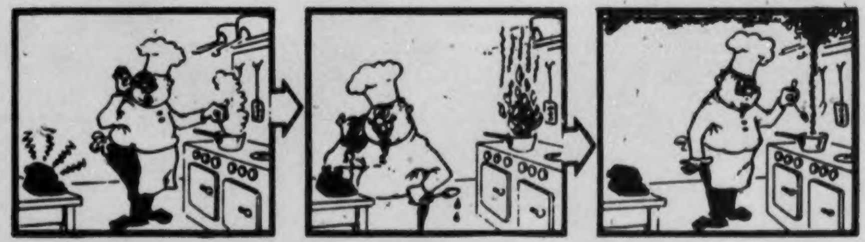
Scoring was based on a series of 24 points concerning particulars of pronunciation (including stress and intonation); a score of 0 or 1 was given for each point depending upon its acceptability. Thus, the total scores could range from 0 to 24.

Fluency. This was always the last-administered Speaking test. As in the case of Structural Control, two forms were devised, one for Population I, and one for the remaining populations. The basic plan was to present pictorial material that the student was asked to describe or otherwise freely respond to, in French. He was expected to produce a series of sentences, perhaps three or four, or more. The test can thus be characterized as a controlled, spoken composition.

In the case of Population I, the Fluency test involved the presentation of two line-drawing pictures (one, of four children having a picnic under a tree, and the other, a boy reading a book and a girl looking out of a window). The student was told to look at these pictures and choose one of the two to describe in French. (No information is available concerning which of the two pictures was chosen more frequently.) He was also told that he could say anything he liked about the picture, provided that it was in French, and that he would have a minute to think about what he was going to say. After the minute was up, the student was told to start speaking; the recorder was switched on, and exactly 30 seconds were allowed for his response.

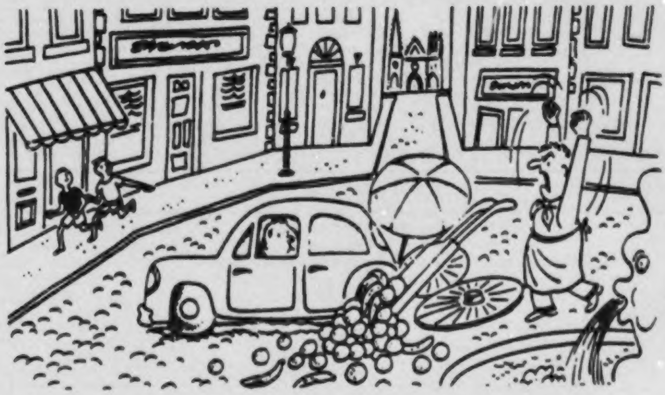
In the case of the test devised for Populations II and IV (and IVS, if there was need), there were two sections similar to the above. In the first section (Fluency 1), the student was presented with three sets of pictures; each set contained three pictures depicting a sequence of events. One of the sets is shown here:

2



The student was asked to inspect the three panels and select one on the basis of which he would give a spoken narration, in French, of what happened. He was then given one minute of preparation time, after which his response (30 seconds maximum) was recorded. In the second section (Fluency 2), the procedure was similar, except that the three stimulus pictures presented for the student's choice were each fairly large and detailed line-drawings of single scenes. One is shown here:

2



The student was told that he "must relate the probable action that took place before the events depicted in the picture, the actual events, and the probable outcome of the situation." (In this way it was hoped that the student would use various tenses.) Again, one minute of preparation time was given, after which any response that took place during a 30-second interval was recorded.

The procedures for the scoring of the French Fluency test were worked out provisionally on the basis of pilot testing using small samples of students of French in England (see Carroll, 1970). The basic rationale for the scoring system was that a weighted sum of a set of fairly detailed, quasi-objective evaluations of specific portions or aspects of the responses would be more valid measures than any "global" rating of the total. Nevertheless, a "global" rating was used as a criterion by which to decide what portions or aspects of the responses would be evaluated, and also to decide on the system of weighting those evaluations to obtain a total score. The data from the pilot testing were sufficient only to suggest the directions in which the development of the scoring system should move. Therefore, the final scoring system was not settled upon until a large number of student response tapes from the final testing had been evaluated according to procedures worked out in the pilot testing.

The analysis of pilot testing data had suggested that the scoring procedures should reflect two somewhat independent aspects of performance: (1) *quantity* of correct response, and (2) relative *quality* of response (essentially, the proportion of responses that were correct). The actual scores produced by the evaluator, listening to the tape-recorded response for each of the one or two sections of the Speaking Fluency test, were as follows:

a=Number of intelligible clauses ("*propositions*," in French; either an independent or a dependent clause; scores [b] through [f] were based only on these "intelligible" clauses);
b=Number of different grammatical structures represented in the response;
c=Number of clauses correct in structure;
d=Number of clauses correct in morphology (inflection);
e=Number of clauses correct in vocabulary;
f=Number of clauses correct in pronunciation.[4]

In addition, after the rater assigned scores to both sections of the Fluency test, a "global rating" using the following scale was assigned to the total response:

0=*très mauvais ou nul* ("very bad, or none at all");
1=*médiocre* ("poor but passable");
2=*honorable* ("satisfactory");
3=*bon* ("good");
4=*très bon* ("very good").

Using data from 1933 cases from various countries and populations, the writer determined scoring formulas for each section of the fluency response as follows:

Fluency Quantity Score=$2b+2c+d+2e+2f$;
Fluency Quality Score=$[6c+3d+6e+6f]/a$;
Total Fluency=Fluency Quantity Score+Fluency Quality Score,
where a, b, c, d, e, and f are as defined above, and scores were set to zero when variable $a=0$.

Where there were two sections of the Speaking Fluency test (as in Populations II and IV), these scores could be summed across the two sections. The formulas were derived in such a way as to predict, more or less optimally, the "global ratings" that were assigned. For the 1933 cases used in the analysis, the Total Quantity Score was correlated with the Total Quality Score to the extent of .728; correlations of Total Quantity Score, Total Quality Score, and Total Fluency (based on two sections of the test) with the global rating were .799, .692, and .814, respectively.

It will be noted that variables b (no. of different grammatical structures), c (no. of clauses correct in structure), d (no. of clauses correct in vocabulary), and f (no. of clauses correct in pronunciation) were about equally important in producing a favorable impression in the rater, with variable e (no. of clauses correct in morphology) of somewhat less importance. The *proportions* of clauses correct in structure, vocabulary, and pronunciation were similarly about equally important.

A scoring formula was also devised for determining a Total Speaking Score. The correlations of the Pronunciation, Structural Control, and Oral Reading sections of the test with the global ratings were relatively low: .399, .651, and .553, respectively. Their correlations with the Total Fluency Scores were .483, .695, and .603, respectively. As a somewhat arbitrary matter, the formula for Total Speaking Score was established as follows:

Total Speaking Score=Total Fluency Score+$2(p+s+o)$,
where p=pronunciation score,
s=structural control score,
o=oral reading score.

As a result, the correlations of the components of Total Speaking Score with that score were as follows: (N=1933)

Pronunciation	.644
Structural Control	.842
Oral Reading	.745
Fluency Quantity	.932
Fluency Quality	.789
Total Fluency	.946

These correlations may be underestimates of the correlations in the complete population because for certain statistical reasons the sample used in this analysis did not include cases with no intelligible clauses.

Writing Tests

As in the case of the Speaking test, the tests of Writing ability were intended for use only with relatively small samples of cases because of the scoring problems involved. Some countries did, however, administer the Writing tests (or certain subtests thereof) to substantial portions of their samples. It was thought highly desirable to obtain evidence on students' abilities to control active production of written French, at least from Population II on up. Writing ability was found not be be an important objective in the teaching of French at the Population I level.

Two different forms of Writing tests were devised, one for Population II and one for Population IV. Essentially, these tests were composed of two parts, one a part that could be reliably scored at national centers, and the other a "directed composition" requiring central scoring by a team of native French speakers (the same team that scored the Speaking tests). These two parts were called Writing I and Writing II, respectively. Writing I had two types of items:

Writing I: Grammar. This consisted of 25 (Population II) or 26 (Population IV) sentences in which particular words (usually short "function" words) were replaced by blanks of a standard size; the student was to write the single missing word on the basis of the context. The sentences used in the Population IV test were in most cases different from those used in the Population II test. Following are Items 13 from the two tests (with the keyed answers):

(Pop. II) *Si j' [étais] riche, j'achèterais une belle voiture.*
(Pop. IV) *Voilà Paul qui arrive; je vais [lui] dire bonjour.*

Writing I: Verbs and Modifiers. This consisted of 7 (Population II) or 6 (Population IV) sentences with certain verbs and modifiers (determiners and adjectives) underlined; each such sentence was accompanied by a parallel sentence of the same form with a change of gender, person, or number, but with a series of blanks; the student was to fill in the blanks to produce grammatical agreement. Examples:

(Pop. II, Item 26) *Mes petits frères font leurs devoirs.*
Ma [petite] soeur [fait] ses devoirs.
(Pop. IV, Item 32) *Vous croyez que vous pourrez y aller tout seul?*
Tu [crois] que tu [pourras] y aller tout seul?

A score of one point for each word written was given if it matched the key exactly. The time limit for Writing I as a whole was 15 minutes.

Writing II: Directed Composition. The requirement set for Population II was to write "a six-exchange dialogue between Louis and Paul, using the 'tu' form throughout." A list of nine words or phrases (*belle journée, promenade,* etc.) was presented, to be used in the dialogue in the order in which they were given, with appropriate grammatical changes, and any additional lexical or structural elements necessary to "tell the story clearly." The test provided spaces for writing the responses; each exchange was to have at least three words. A period of 10 minutes was allowed.

For Population IV the requirement was to "write a short composition comparing the merits of living in the country and in a big city." Several "themes" were suggested (in French); for country life: *tranquillité, paysages, nourriture, santé;* for city life: *distractions et amusements, confort moderne, grands magasins, études supérieures.* The student had 25 minutes to write his composition, taking all the space he might need.

The scoring system for the directed compositions was developed in a manner similar to that for the Speaking Fluency test (for statistical analysis of pilot testing data, see Carroll, 1970). Again, it was felt desirable to compute separate scores for *quantity* and *quality* of writing. The final scores were computed on the basis of the following variables:

a=Number of intelligible clauses;
b=Number of clauses which are grammatically correct;
c=Number of clauses which are correct in vocabulary;

d=Proportion of clauses which are grammatically correct=b/a;
e=Proportion of clauses which are correct in vocabulary=c/a.

From these, the final scores were computed as follows:

Writing Quantity=$3b+2c$;
Writing Quality=$40d+30e$;
Total Writing II=Writing Quantity+Writing Quality.

The integral weights included in these formulas were arrived at through regression analysis such that there would be nearly optimal prediction of global ratings that had been made of the responses on a scale of 1 to 5, 5 being the high end of the scale. (The scale points were the same as those for the global ratings of the Speaking Fluency responses, except for the systematic translation of the scale upward by one point.) These regression analyses were based on 13 separate samples from various countries and populations, and from the pooled samples, the total N being 11,889. (Three hundred ninety-eight cases were not used because they had zero scores on "number of intelligible clauses," or were defective in other ways.) The results clearly supported the hypothesis that the quantity and quality aspects were somewhat independent and should be scored separately. The final Quantity and Quality scores correlated .712; their correlations with global ratings were .710 and .668, respectively. (For further data concerning intercorrelations of the various scores, computed for the complete samples in the final testing, see below.)

No formula was devised for combining Writing I and Writing II scores; these scores were considered separately in all data analyses, largely because the N's involved were generally quite different. Writing I was, after all, primarily a measure of active mastery of grammar, whereas Writing II was a measure of much more than grammar.

TECHNICAL CHARACTERISTICS OF THE PROFICIENCY TESTS

This section presents a discussion intended to help the reader interpret scores on the proficiency tests in terms of what they measure, what levels of proficiency are represented by particular scores, and how accurate the measurements are. Several other issues of interest will be commented on; for example, the question of whether the tests are equally effective for native speakers of different languages.

Throughout this discussion, it is necessary to present concrete results of the testing from the several countries and populations. The reader is cautioned, however, that he should resist the temptation to draw conclusions regarding the comparative achievements of the samples. It is beyond the scope of this chapter to attempt explanations of differential performances; discussion of the testing results from this standpoint is undertaken only in fuller technical reports.[5]

Score Distributions

Test score distributions are a function both of the characteristics of the tests and of the nature of the samples that are tested. In an assessment project like the IEA study it is considered desirable to have tests that spread out the members of the populations tested from relatively poor performance to relatively excellent performance, and at the same time exhibit the relative numbers of the samples performing at different levels. The purpose of this section is to examine the extent to which this was true for the various tests of proficiency in French that were used in the IEA study of French as a Foreign Language.

Because of the voluminous data collected, only the results for populations from the United States will be presented--not because they are typical, but because they are generally from large samples, and because data are available for all tests and from all populations. The test score scales that will be shown, however, have ranges on which the results for all other countries can be projected.

The test score distributions have been extracted from computer tabulations in which the scores were classified into anywhere from 4 to 13 class intervals; in some cases (namely, those for formula-corrected scores), these class intervals were coded by the computer program into positive integers by a linear transform from the original scores. In all cases, the distributions shown here are in terms of the original score scales; i.e., coded scores have been transformed back into the original scales. Histograms, where shown, are drawn to include the true (generally, fractional) lower and upper bounds of class intervals. (For example, a class interval coded 3 and covering rounded integral scores of 3, 4, and 5 is considered to have true lower and upper bounds of 2.5 and 5.5, respectively.) Frequency polygons, where shown, are drawn so that the data points are at midpoints of class intervals. (For example, the midpoint of the class interval just mentioned would be 4.) While grouping of scores loses some precision, it serves to give a more accurate impression of the general shape of the distributions.

Listening Tests. Relevant score distributions are shown in Figure 1: (a) for the sample (*N*=206) in USA Population I tested with the test appropriate for that population; (b) for the 19 items that were common to the Population II and IV tests, displaying relative performance of Populations II (*N*=1924) and IV (*N*=1370) on these items; and (c) for the Population II and IV tests and corresponding samples (*N*=1927 and 1370, respectively) projected onto a common scale (see the section below on the equating of these tests). The Population I test yielded a more or less normal ("bell-shaped") distribution ranging from chance-level scores around 0 up to perfect scores; it was somewhat too easy for around 25 percent of the sample, however, who got scores at or close to the maximum. The Population II/IV anchor items effectively separated the two populations, as one would expect, yet gave wide variation of performance for both populations. The complete Population II and IV Listening tests were also of appropriate difficulty for the respective samples and yielded wide variations in level of performance.

Just as it is difficult to express level of competence in understanding French in any concrete terms, it is difficult to interpret levels of test performance in those terms. Considering the level of difficulty (and speaking rate) of the materials in the Population IV test, we may assume, however, that top performers on that test (of which there were relatively few in the USA Population IV) would be able quite easily to understand everyday conversations and typical broadcast announcements in French. The Population II test, however, is probably not difficult enough to assess this level of understanding, and the Population I test, even less so.

Reading Tests. Relevant score distributions are shown in Figure 2, analogously to Figure 1. The Population I test, designed to be very easy, is actually too easy for a few members of that population in the United States, but does show wide variation in levels of performance. The anchor items for the Population II and IV tests are somewhat too hard for Population II members, and somewhat too easy for members of Population IV. The complete Population II Reading test is difficult for the USA Population II sample, but effectively spreads out the better members of the sample; the complete Population IV Reading test yields effective differentiation of performance levels for USA members of that population. Top-level performance on the Population IV test probably corresponds to the ability to read, with reasonable understanding, written French of the type found in popular novels and newspapers. Thus we may assume that in the USA Population IV, the average level of competence in reading French falls considerably short of that level of ability. The average level of competence in the USA Population II is much less than that of Population IV; most members have difficulty in reading even very simple French.

Speaking Tests. Speaking tests were given to relatively small numbers of cases in the respective populations. Figure 3 shows distributions for three subtests that were given before the Fluency test. In Figure 3*a*, the distributions are for precisely the same test given to all three populations; there is much variance both within and among these populations. The average performance even in Population IV falls considerably short of acceptable pronunciation of the 29 pronunciation details that were assessed (one point being given for each detail correctly pronounced). Presumably, native French speakers would get perfect or near perfect scores. Figure 3*b* shows distributions for Structural Control; it should be noted, however, that the Population I test was designed to be easier than the test designed for Populations II and IV. Each test consisted of ten items evaluated on a scale ranging from 0 to

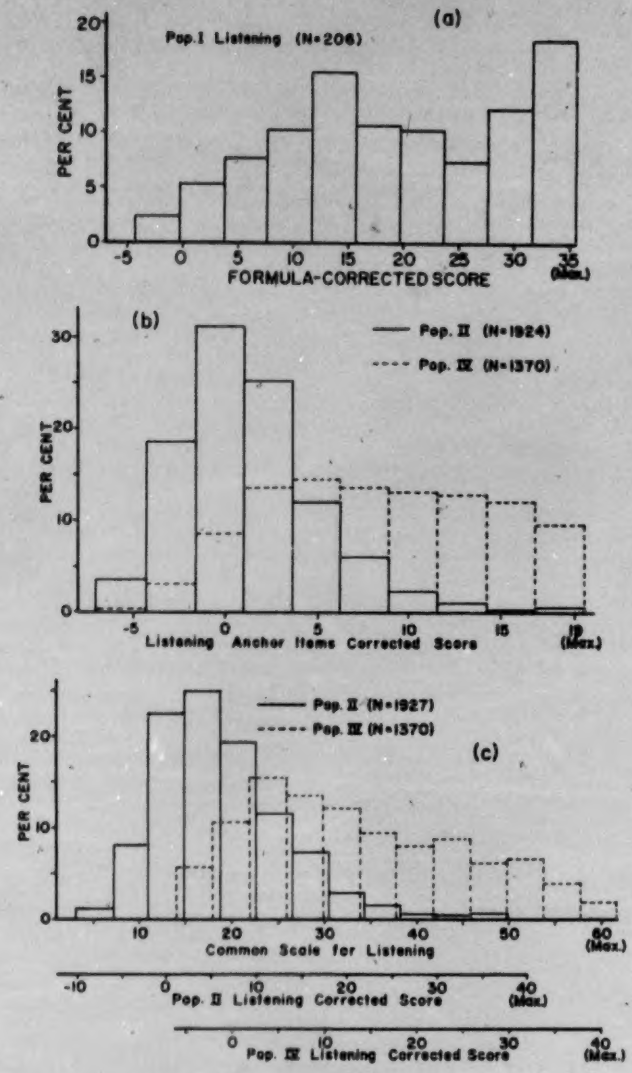


Figure 1. Distributions of scores (by class-intervals) for Listening tests in the USA samples: (*a*) for the Population I test; (*b*) for the anchor items common to the Population II and IV tests; (*c*) for the complete Population II and IV tests on a common scale determined by vertical equating procedures.

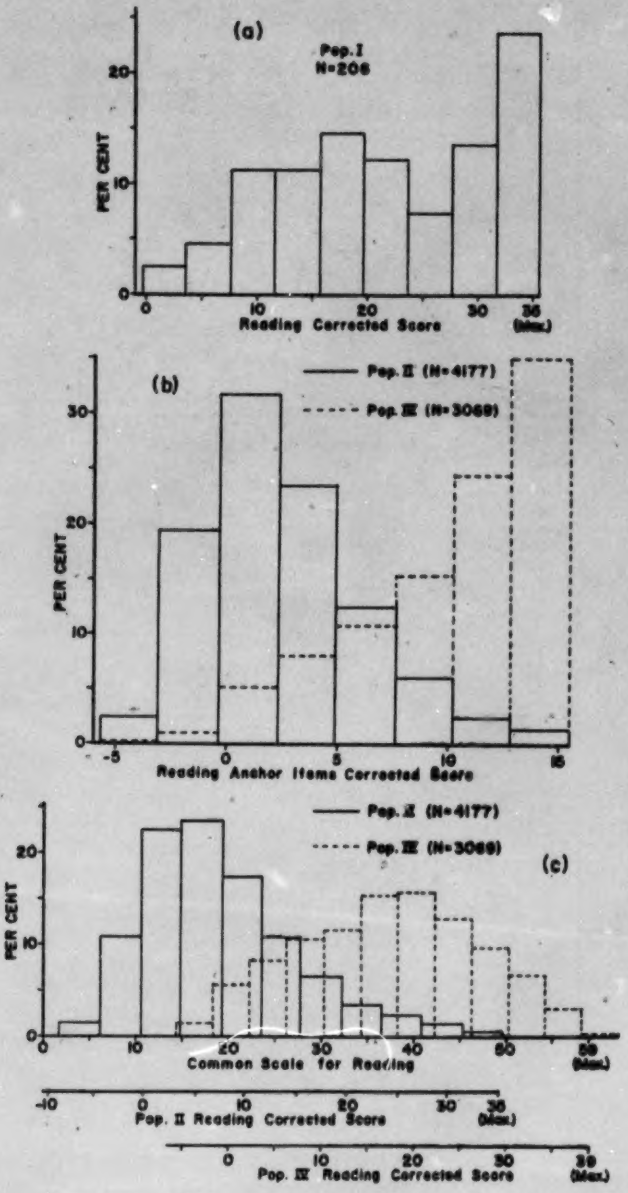


Figure 2. Distributions of scores (by class-intervals) for Reading tests in the USA samples: (*a*) for the Population I test; (*b*) for the anchor items common to the Population II and IV tests; (*c*) for the complete Population II and IV tests on a common scale determined by vertical equating procedures.

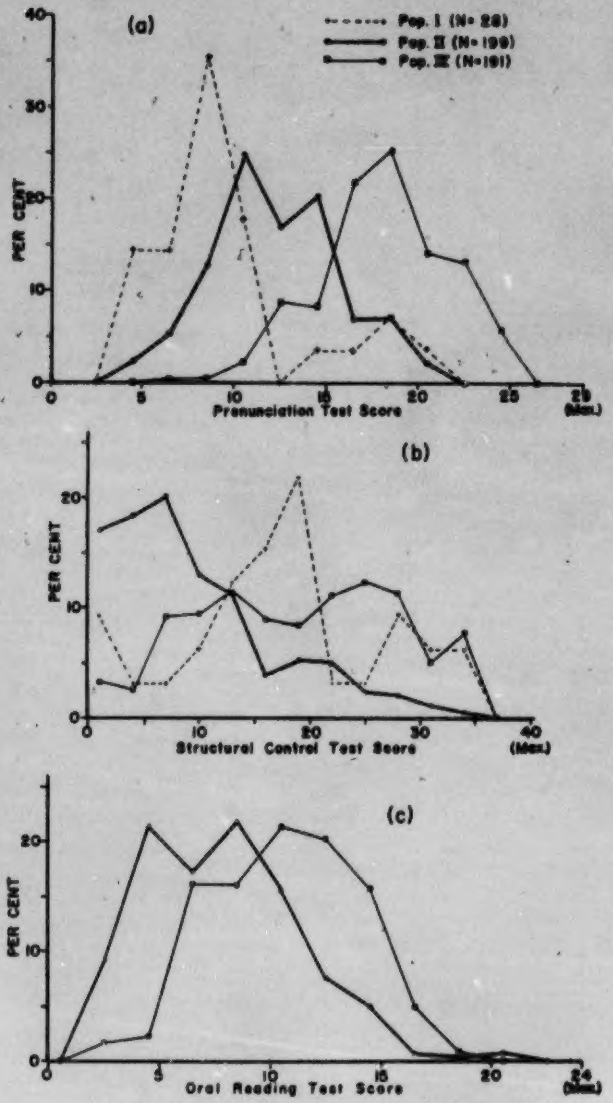


Figure 3. Distributions of scores (by class-intervals) for USA samples on: (*a*) the Pronunciation test (the same test for all populations); (*b*) the Structural Control tests (the test for Population I is different from that for Populations II and IV); (*c*) the Oral Reading test (the same test for both Populations II and IV).

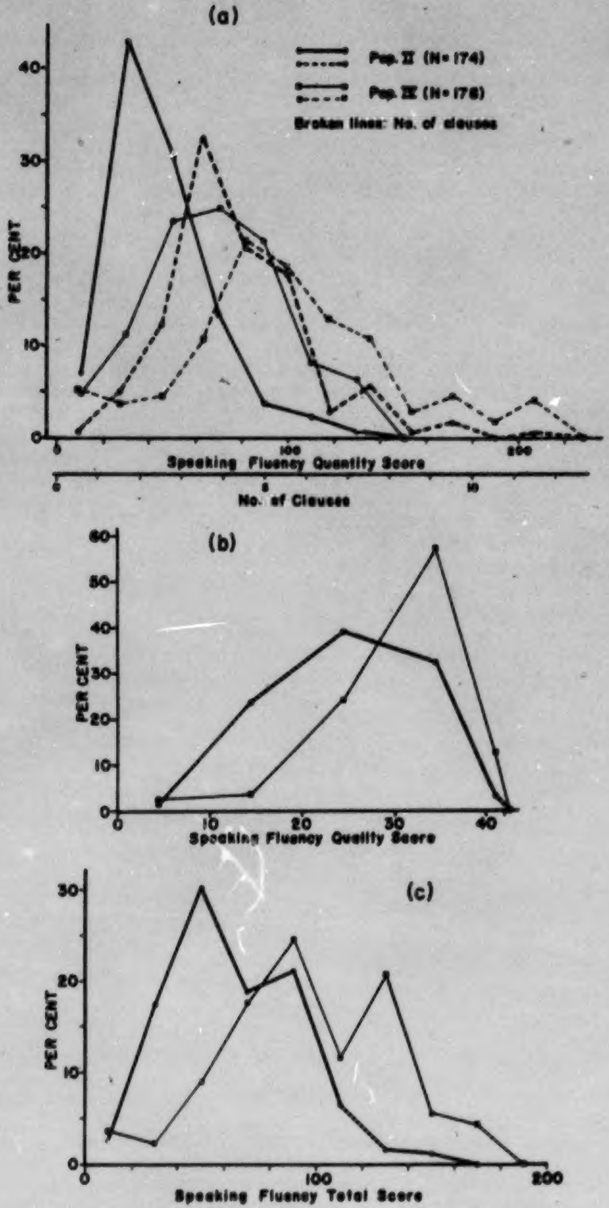


Figure 4. Distributions of scores (by class-intervals) for USA samples on the Speaking Fluency test (the same test for Populations II and IV): (*a*) Quantity score; (*b*) Quality score; (*c*) Total Fluency score. In Figure 4(*a*) the broken lines are for distributions of numbers of "intelligible clauses" in the responses to the second part of the Fluency test, scaled to correspond to 18 points per clause on the Quantity score scale. Distributions of scores for Population I are not available, but the mean Quantity score for the sample (*N*=28) from this population was 23.3 with a standard deviation of 16.7 (the scores having been multiplied by a factor of 2 to make them comparable with the Population II and IV scale).

4; thus, scale points of 0, 10, 20, 30, and 40 can be interpreted as representing average ratings at corresponding points on the 0 to 4 scale. (See page 18 for a specification of the scale points.) There is a wide spread of performance level in all populations. Since the problems were relatively easy, even a top score would probably not indicate complete "structural control" of spoken French. Figure 3*c* shows distributions of scores on an Oral Reading test that was identical for Populations II and IV (it was not given to Population I). It was scored like the Pronunciation test, i.e., a point was given for each of 24 pronunciation details. The distributions of scores in the USA samples were more or less normal.

Distributions of the scores derived from the Fluency test given to members of Populations II and IV are given in Figure 4. (Distributions of the scores for the Population I test are not available.) In Figure 4*a*, the solid lines are frequency polygons for the "Quantity" scores for Populations II and IV. These scores might more appropriately be called "quantity *correct*" scores, since they were derived from a weighted scoring of clauses for several aspects of correctness or quality. Fundamentally, however, they depend on how many clauses the subject gave in a 30-second period. In view of the weighting system, it can be said that for each clause that was "different in structure" and correct in structure, morphology, and vocabulary, the subject would receive nine points. Because there were two parts to the Fluency test, and distributions are available only for the separate parts, to provide a basis of comparison with the Quantity score an additional scale is given on the baseline of Figure 4*a* indicating 18 points for each clause, and the broken lines show frequency polygons for the number of "intelligible clauses" (regardless of correctness) in the responses to the second part of the Fluency test. (These distributions are generally similar to those for the first part of the Fluency test.)

Comparison of the sheer quantity of responses with the "quantity correct" scores is reflected in the "Quality" scores whose distributions are shown in Figure 4*b*. According to the system of weighting points for acceptability of responses clause by clause, a subject whose response (regardless of the number of clauses, as long as there was at least one) was completely acceptable would receive a "Quality" score of 42. From the distributions shown in Figure 4*b*, it may be said that the median quality for Population II was to get about half the maximum possible number of points for correctness, while Population IV students typically received about 70 to 80 percent of the maximum possible number of points. There were, of course, wide individual differences.

The Total Fluency scores, whose distributions are shown in Figure 4*c*, are the simple sums of Quantity and Quality scores; there is no maximum possible score. Scores range widely, and Population IV does significantly better than Population II. It would be of interest to give the test to native speakers of French to ascertain their typical level of performance; probably it would be somewhat above the top scores attained by members of Population IV (assuming that native French speakers would be talkative enough to produce something like at least 10 intelligible clauses in a 30-second period).

The distributions of the "global ratings" given to the fluency responses, shown in Figure 5*a*, may give an impression of the overall acceptability of these spoken responses. The distribution for 28 cases in Population I is included here; it is only in that population that any significant proportion of responses were given top ratings ("4, *très bon*")--possibly the raters were more indulgent in evaluating 10-year-olds' responses. At any rate, substantial proportions of Population IV responses were awarded 3, "*bon*." The relative ease of the Fluency test tasks, in comparison to normal conversational speaking requirements, must of course be considered in assessing what the results mean.

The distributions of Total Speaking scores (weighted composites of all Speaking test scores) are shown in Figure 5*b*. There is no absolute maximum possible score, but one may suppose that native French speakers would obtain scores somewhat above the top performance represented in the distributions.

Writing Tests. Figure 6 shows distributions of scores on the tests titled Writing I, designed for Populations II and IV. These are tests of grammatical knowledge; although some items are in common between the tests, they are at different levels of difficulty and have different maximum possible scores. (Depiction of

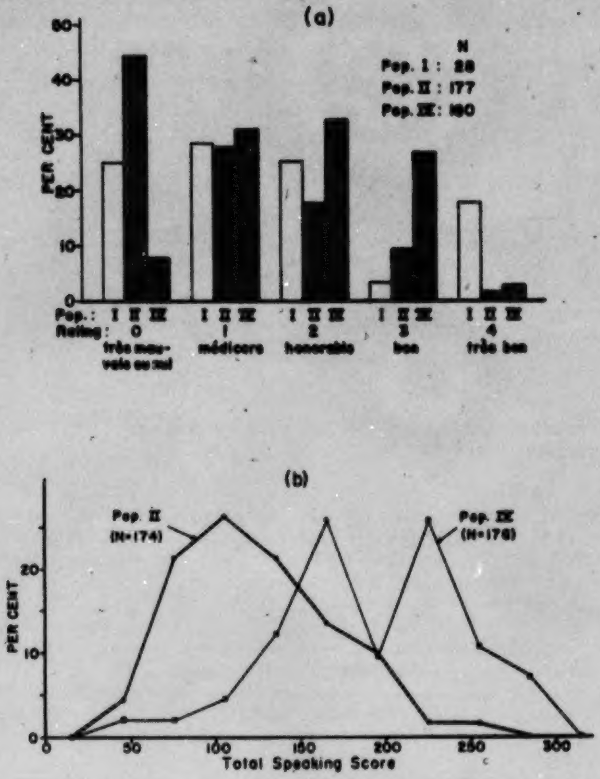


Figure 5. Distribution of (*a*) global ratings of Speaking Fluency responses, and (*b*) Total Speaking scores (by class-intervals) for USA samples.

the distributions on the same scale is done only for convenience.) Both tests tend to be difficult for the USA populations, as is evident from the positively skewed distributions.

Distributions of scores derived from Writing II (Fluency) are shown in Figure 7. In Figure 7*a*, the solid lines are frequency polygons for the "Quantity" (actually, quantity correct) scores, USA Populations II and IV. Interpretation of these scores follows somewhat the same logic as the quantity scores from the Speaking Fluency test, since they depend partly on the number of clauses found in the student's response (for which 10 minutes was allowed in the Population II test, and 25 minutes or more in the Population IV test, the tests themselves being quite different). In view of the weighting system, it can be said that for each clause that was correct in grammar and vocabulary, the subject would receive 5 points. The broken lines in Figure 7*a* (with the corresponding scale on the baseline) show distributions of the sheer numbers of "intelligible clauses" in the responses, regardless of correctness. Population II students typically wrote about 7 such clauses in 10 minutes, while Population IV students typically wrote about 13 clauses in 25 minutes. (A few

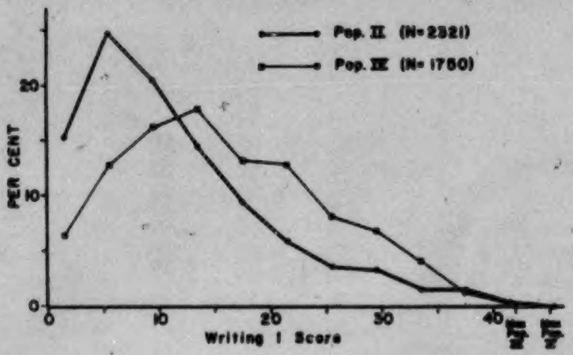


Figure 6. Distributions of scores (by class-intervals) of USA samples on the Writing I tests (different tests for Populations II and IV).

students wrote up to some 30 clauses.) But the typical student in Population II had only 3 clauses correct in both grammar and vocabulary, and the typical student in Population IV had about 5 or 6 clauses thus correct.

As in the case of the Speaking Fluency test, comparison of the sheer quantity of responses with the Quantity scores is reflected in the "Quality" scores whose distributions are shown in Figure 7*b*. According to the weighting system, a subject whose response (regardless of length, provided that there was at least one clause) was completely acceptable in both grammar and vocabulary would receive a "Quality" score of 70. As may be seen from Figure 7*b*, it was the case in both Populations II and IV that the typical student received about half the number of points possible for the number of clauses in his composition. A very small number of students received perfect scores.

The Total Writing Fluency scores were obtained by simply summing the Quantity and Quality scores; the distributions are shown in Figure 7*c*. It should be remembered that the Population II and IV tests were different; the distributions of scores are therefore not directly comparable, even though for convenience they are plotted on the same scale.

Figure 8 shows distributions of global ratings assigned to the compositions; here the distributions for Populations II and IV are distinctly different. No student in either population received a top rating ("5, *très bon*"). One wonders, of course, what distributions of ratings would have been assigned to French compositions written by samples of native French speakers with comparable ages and degrees of education.

Reliability

Listening Tests. Reliability coefficients for *raw scores* on the Listening tests were computed using the Kuder-Richardson Formula 20 (Kuder and Richardson, 1937; also see Lord and Novick, 1968, Chapter 4). This formula utilizes internal consistency data and the item difficulties. Separate coefficients were computed for each test in each country. It is well known, however, that the reliability coefficient is a function of the variance of total scores, but that the standard error of measurement (under certain assumptions) should be constant despite variation in variances of total scores. The lower portion of Table 5 shows means, standard deviations, reliabilities, and standard errors of measurement for each test in each country where data were available. As may be seen, the standard error of measurement for the Population I Listening test (given only in the United States)

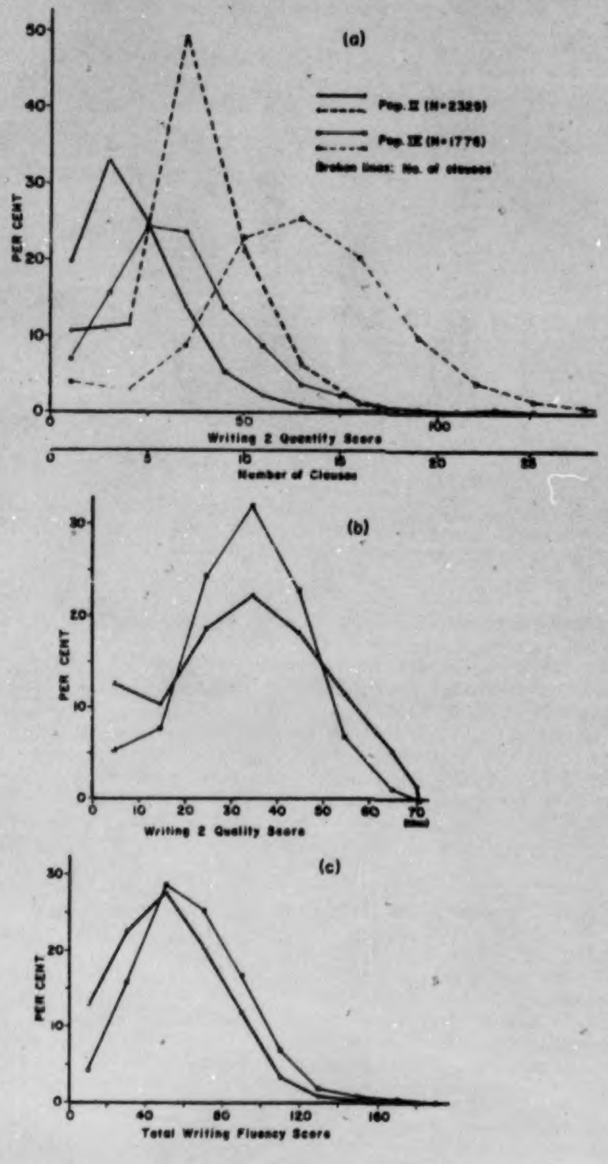


Figure 7. Distributions of scores (by class-intervals) for USA samples on the Writing II tests (different tests for Populations II and IV): (*a*) Quantity score; (*b*) Quality score; (*c*) Total Writing Fluency score. In (*a*) the broken lines are for distributions of numbers of "intelligible clauses" in the responses, scaled to correspond to 5 points per clause on the Quantity score scale.

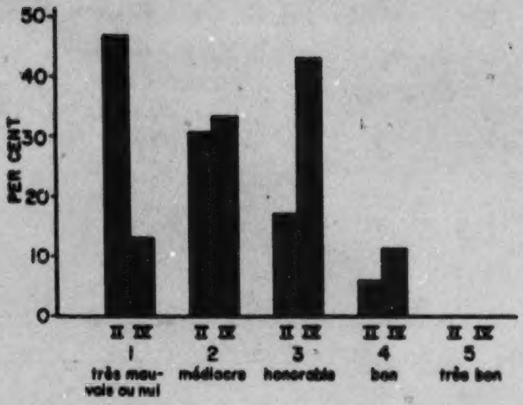


Figure 8. Distributions of global ratings for USA samples on the Writing II tests (different tests for Populations II and IV).

is 2.37, where the total possible score range is 0 to 35. The standard errors of measurement for the Population II Listening test (with a possible score range of 0 to 40) are all in a narrow range from 2.72 to 2.78 except for the value 2.15, for Rumania. This latter value may have resulted from the fact that Rumania's mean is quite high, and the score distribution is somewhat negatively skewed (suggesting that the top score is a "ceiling" that fails to take account of further variation in the upper range of competence). The standard errors of measurement for the Population IV Listening test (score range 0-40) seem to be a curvilinear function of mean scores, with values of 2.43 for Chile's relatively low mean score, values in the range 2.60 to 2.80 for countries in the middle range of mean scores, and values of 1.91 and 2.21 for two countries with relatively high mean scores. A curvilinear relation between mean score and standard error of measurement may be expected according to test score theory.

While these standard errors of measurement may not be considered sufficiently small to justify using the tests for comparing individuals with closely similar scores, they are certainly small enough to justify using the tests for comparing group means, and for other research purposes.

Unfortunately, the computations that have been made available do not permit estimating the reliabilities or standard errors of measurement for formula-corrected scores. It may be expected, however, that such statistics should be more or less comparable to those for raw scores.

Reading Tests. Precisely the same procedures were followed in computing reliability coefficients and standard errors of measurement for Reading test raw scores; these are shown in the upper portion of Table 5. Again, no statistics on reliability were available for formula-corrected scores.

The reader may inspect the data and see that the standard errors vary in much the same manner, and lead to conclusions and remarks that are similar to those made for the Listening test.

Speaking Tests. No precise data are available concerning the reliabilities or standard errors of measurement of the Pronunciation, Structural Control, or Oral Reading tests, but lower bounds for the reliabilities can be computed from distribution statistics (means and standard deviations) by using Kuder-Richardson Formula 21 (Lord and Novick, Chapter 4) and the corresponding standard errors of measurement can then be derived. On this basis, the reliabilities for the Pronunciation score varied widely (from .029 to .774; median, .657), but the standard errors of

Table 5

Means, Standard Deviations, Reliability Coefficients (KR 20), and Standard Errors of Measurement for Reading and Listening Test Raw Scores

	Population I and II*				Population IV			
	M	SD	Rel.	SE	M	SD	Rel.	SE
Reading Tests								
USA I	25.1	7.2	.897	2.31				
Chile	--	--	--	--	13.3	4.8	.684	2.70
England	15.3	7.2	.873	2.57	33.8	4.0	.789	1.84
Netherlands	18.0	6.0	.814	2.59	29.2	4.8	.768	2.31
New Zealand	17.6	6.3	.832	2.58	30.1	4.2	.725	2.20
Rumania	28.8	6.6	.918	1.89	31.0	6.8	.906	2.02
Scotland	18.3	7.5	.887	2.52	28.4	5.5	.829	2.28
Sweden	--	--	--	--	24.2	7.5	.883	2.57
USA	14.2	5.7	.788	2.63	22.5	7.3	.877	2.56
Listening Tests								
USA I	23.3	8.0	.912	2.37				
Chile	--	--	--	--	10.4	4.8	.744	2.43
England	17.0	7.5	.866	2.74	30.7	6.0	.864	2.21
Netherlands	--	--	--	--	--	--	--	--
New Zealand	13.3	4.7	.665	2.72	19.4	6.3	.803	2.80
Rumania	26.7	12.2	.969	2.15	28.1	12.3	.976	1.91
Scotland	16.5	6.4	.812	2.78	22.8	6.8	.843	2.70
Sweden	--	--	--	--	24.0	7.9	.892	2.60
USA	14.4	5.5	.752	2.74	20.1	8.7	.905	2.68

*Population II except for USA I

measurement for different countries and populations all fell within the range 2.41 to 2.66 (median, 2.55), on a score scale from 0 to 29. For Structural Control, reliabilities were generally high (.557 to .916; median, .862), with the standard errors of measurement being within the range 2.30 to 2.99 (median, 2.78), on a score scale from 0 to 40. Reliabilities for Oral Reading ranged from .067 to .708 (median, .553), with the standard errors of measurement ranging from 2.01 to 2.42 (median, 2.33), on a score scale of 0 to 24. These standard errors of measurement are all sufficiently small, in relation to the score scale ranges, to justify use of these variables in group comparisons.

A special experiment was conducted to estimate the inter-rater reliabilities of various scores and ratings from the Speaking Fluency tests. Data for 63 cases selected to give a wide dispersion of apparent quality were evaluated by two different raters (different pairs of raters from the team of eight). The intra-class correlation between the ratings was .737 for the global Fluency rating. This is a quite respectable figure; reliabilities of subjective global ratings are often considerably lower.[6] The inter-rater reliabilities of various detailed ratings and composites thereof varied considerably. The highest reliabilities (intra-class correlations) occur for variables in which raters count clauses (.983, .976 for "number of intelligible clauses" in the two sections of the Fluency test), or count clauses that are correct in certain respects (reliabilities around .85). These relatively high reliabilities are reflected in those for Fluency Quantity scores (reliabilities of about .95). Raters do not agree so well when they are counting numbers of different structures (reliabilities of .69 and .62 for two sections of the test). When the Quantity scores are converted to proportions (in computing Quality scores), the reliabilities are reduced considerably (reliabilities around .77). Nevertheless, the final Total Fluency Score has an inter-rater reliability of .934.

It is well known, and obvious in the nature of things, that inter-rater reliabilities are overestimates of the reliabilities of measurements from the standpoint of test form equivalence, or repeated measurements. Because the Fluency test contained two sections (A and B) that might be considered equivalent forms or repeated measurements, it was possible to compile some limited statistics on the repeated-measurement reliability of the Fluency measurements. Statistics based on the total sample of 1933 cases from both Populations II and IV (pooled from several countries) used in the analysis leading to composite score weighting formulas are shown in Table 6; also shown are the reliabilities of the total scores (summing over the two Fluency sections) estimated by the well-known Spearman-Brown Formula. These

Table 6

Inter-Form Reliabilities, Speaking Test (Fluency A vs. Fluency B)

N=1933

Measure	r_{AB}	Estimated reliability of Total Score*
No. of Intelligible Clauses	.628	.771
No. of Different Structures	.548	.708
No. of Clauses Correct in Structure	.552	.711
No. of Clauses Correct in Morphology	.575	.730
No. of Clauses Correct in Vocabulary	.552	.711
No. of Clauses Correct in Pronunciation	.638	.779
Quantity Score	.671	.803
Quality Score	.527	.690
Total Fluency Score	.666	.800

*Estimated from r_{AB} by the Spearman-Brown Formula (Lord and Novick, 1968, p. 84).

reliabilities reflect the consistency with which a student demonstrated his competence on the two sections of the Fluency test, but of course they are somewhat attenuated by the less-than-perfect inter-rater reliabilities. The estimated repeated-measurement reliability of the Total Fluency Score, .800, may be considered to indicate very satisfactory reliability for this type of test. Its magnitude partly reflects the successful use of semi-objective procedures in evaluating the responses.

We have no way of estimating the reliability of the Total Speaking Score, but can estimate it as being quite high in view of the relatively high reliabilities of its components.

Writing Tests. Lower bounds of reliabilities of Writing I scores were estimated by use of distribution statistics in Kuder-Richardson Formula 21. These are shown in Table 7 and are in most cases quite high.

Table 7

*Means, Standard Deviations, Reliabilities, and Standard Errors of Measurement of Writing I Scores**

	Population II				Population IV			
	M	SD	Rel.	SE	M	SD	Rel.	SE
Chile	--	--	--	--	3.5	3.4	.740	1.73
England	10.2	9.8	.939	2.43	31.5	6.5	.834	2.65
New Zealand	13.0	9.1	.909	2.75	23.5	7.7	.846	3.03
Rumania	28.0	11.9	.946	2.76	28.0	11.3	.950	2.54
Scotland	14.0	11.2	.944	2.65	22.2	9.2	.898	2.94
Sweden	--	--	--	--	15.8	8.4	.881	2.89
USA	11.3	8.4	.900	2.66	16.0	8.8	.893	2.87

*Reliabilities estimated by Kuder-Richardson Formula 21

For Writing II (Writing Fluency), investigations of inter-rater reliability followed much the same procedures as were used for Speaking Fluency scores. The global ratings had inter-rater reliabilities of .772 and .748. The two figures given are for Population II (N=70) and Population IV (N=70) cases, respectively. For the variables that entered into final composites, reliabilities were: .966 and .984 (no. of intelligible clauses), .954 and .936 (no. of intelligible clauses with correct grammar), and .939 and .926 (no. of intelligible clauses with correct vocabulary). The resulting inter-rater reliabilities for the Quantity Score were .967 and .955; for the Quality Score, .945 and .902; and for the Total Writing Fluency Score, .959 and .937. Thus, the rating team produced evaluations that were highly consistent from one rater to another.

The data provided no way of estimating repeated-measurements reliabilities to indicate how consistent the students would be in their Writing Fluency performance if they were retested.

Item Analyses

Item analysis tabulations were made only for the Reading and Listening tests; nevertheless, these are voluminous, since they were made for each population sample in each country where the corresponding test was given. The item analyses showed, for each item, the item difficulty (expressed as the proportion of examinees answering correctly), the point-biserial correlation of the item with a total score based on the *remaining* items, the response distribution (proportions of examinees selecting each possible response, giving no response, or not reaching the item) and point-biserial correlations for each possible response.

Only samples of results from these item analyses will be presented here, because of space limitations. Table 8 shows distributions of item difficulties for certain countries and populations; item difficulties are expressed as "deltas"; to obtain a value of delta, the proportion failing is converted to its corresponding normal deviate (x), from which delta is computed as $\Delta=13+4x$. Thus, an item failed by 50 percent of the examinees in a particular sample receives $\Delta=13.0$; an item failed by only 15.87 percent (an easy item) corresponds to $x=-1.0$ and $\Delta=9.0$. The mean values of delta ($\bar{\Delta}$) are inversely related to mean raw scores. The standard deviation of Δ

Table 8

Distributions of Deltas for Selected Countries and Populations on French Reading and Listening Tests

	USA I		Rum. II		USA II		Chile		Rum. IV		USA IV	
Delta*	R	L	R	L	R	L	R	L	R	L	R	L
20							1	2				
19						1	0	4				
18						3	3	8			1	
17					4	5	8	7		1	0	3
16					4	8	4	7		1	6	3
15				1	13	8	10	8	2	2	3	4
14		4		4	5	7	8	1	1	1	4	9
13	7	6	2	5	5	3	1	0	0	1	6	11
12	7	9	2	11	1	4	3	1	4	10	2	5
11	9	9	7	8	2	1	1	0	5	5	11	3
10	8	6	8	10	1			0	9	15	3	2
9	3	1	6	1				1	7	4	0	
8	0		5					1	4		3	
7	1		3						4			
6			1						2			
5			0						0			
4			1						0			
3									1			
No. items	35	35	35	40	35	40	39	40	39	40	39	40
Mean Delta	10.53	11.20	8.92	11.17	14.01	14.58	14.82	15.84	9.19	10.74	12.09	12.97
S.D. Delta	1.47	1.27	1.87	1.52	1.64	1.86	1.94	2.43	2.38	1.82	2.36	1.72
N	206	206	2271	117	4177	1927	1440	173	2247	77	3069	1370

*See text for explanation of delta. Each class interval represented here ranges from the integral value shown here down to just above the next integral value below; i.e., the class interval labeled "20" ranges from 20.00 down to 19.005.

indicates the variability of the deltas. The distributions of Δ's for the tests, countries, and populations shown in Table 8 reflect what may be regarded as desirable for tests designed to measure a fairly wide range of competence.

Table 9 shows distributions of item validities for certain countries and populations; these are expressed as Fisher's z-transformations of point-biserial correlations (this transformation tends to normalize the distributions). The mean value, $\bar{z}$, is in every case significantly positive, and in many instances it is quite high, particularly where the standard deviation of the raw scores is also high. It was very rare that an item was found to have a negative validity, but this was always in particular countries and samples; such an item generally had significantly positive validities in other countries. The item analysis results contain few, if any, indications that items should be eliminated or further modified in any future editions of the tests. It will be recalled that the tests had gone through several cycles of

Table 9

Distributions of z's (Transformed from Point-Biserial Correlations) for Selected Countries and Populations on French Reading and Listening Tests

	USA I		Rum. II		USA II		Chile IV		Rum. IV		USA IV		r_{pb}
z	R	L	R	L	R	L	R	L	R	L	R	L	At Midpoint
-.10 - -.051								1					-.075
-.05 - -.001					1		1	1					-.025
.00 - .049					0	1	2	4					.025
.05 - .099				1	0	2	4	2					.075
.10 - .149				0	1	6	8	3	2				.124
.15 - .199	1		1	0	2	7	5	6	0		1	1	.173
.20 - .249	0		1	1	7	7	11	6	1		4	4	.221
.25 - .299	2		1	0	9	7	3	2	3		5	2	.268
.30 - .349	5	3	3	0	7	5	3	7	3	3	3	2	.314
.35 - .399	4	3	0	1	6	2	2	3	1	2	7	3	.358
.40 - .449	4	7	2	1	1	2		0	7	0	3	9	.401
.45 - .499	7	6	3	0	1	1		5	6	2	8	4	.442
.50 - .549	6	4	8	2					2	1	5	5	.482
.55 - .599	4	8	6	1					8	0	3	4	.519
.60 - .649	0	2	6	4					3	1		3	.555
.65 - .699	2	2	3	4					2	0		3	.588
.70 - .749			0	2					1	3			.620
.75 - .799			1	2						3			.650
.80 - .99				7						5			.716
1.00 - 1.19				6						5			.800
1.20 - 1.39				3						6			.862
1.40 - 1.59				3						5			.905
1.60 - 1.79				2						2			.935
1.80 - 1.99										2			.956
Mean z	.449	.492	.520	.885	.284	.233	.190	.230	.465	1.015	.392	.447	
Corresponding r_{pb}	.421	.456	.478	.709	.277	.229	.188	.226	.434	.768	.373	.419	

pretesting to produce items with desirable characteristics.

Mother-Tongue Differences. A question of some interest in contrastive linguistics is that of whether the items functioned in the same way for speakers of different languages. Statistical examination of the item analysis data indicates that the answer to this question tends to be in the negative, but that the tendency is relatively slight. There are a few items in which the item difficulty in a particular country deviates somewhat markedly from what would be expected on the basis of (*a*) the mean delta for the item for a number of countries and (*b*) that country's mean delta for all items in the test. On the whole, however, most items seem to function in more or less the same way in all countries, regardless of the mother tongue.

The technique by which this was examined was essentially that of looking at specific item-language interactions, along with computations of average correlations of item-difficulties (deltas), in (1) countries with different mother tongues, and (2) countries with the same mother tongue (English).

If we look at Table 10, we note that the standard deviations of residuals from expectancies (technically, the standard deviations of interaction terms in the basic equation of the analysis of variance) are always much larger (statistically, very significantly so) for a set of countries with different mother tongues than they are for a set of countries with English as the mother tongue. Also, the average intercorrelations of deltas across countries are smaller for sets of countries with different mother tongues. These findings, however, may be traced to the deviant behavior of a relatively small number of items.

For the interest of any reader who may want to follow up this matter, we give here a list of "deviant" item-country combinations, i.e., all those items whose deviation from expectancy, measured on the scale of delta, exceeded 2 points (one-half standard deviation); the figures in parentheses indicate the actual deviation. Thus (+2.42) for Rumania, Population II Reading, means that the Rumanian sample did that much better than expected on that item; (-3.21) for Rumania, Population II Listening test, means that the sample did that much worse than expected. All these deviations occurred in the across-mother-tongue comparisons; there was only one

Table 10

Analysis of Item Analysis Statistics Among and Within Mother-Tongue Groups

Test	Population	Countries Pooled	Deltas from Pooled Countries: Mean	Deltas from Pooled Countries: SD	SD of Residuals	Average Intercorrelation of Residuals
Reading	II(III)*	Neth,Rum,Swe, USA	12.13	3.37	1.00	.721
		Eng,NZ,Scot,USA	13.33	3.48	.52	.917
Listening	II(III)*	Rum,Swe,USA	12.88	3.07	1.29	.611
		Eng,NZ,Scot,USA	14.34	3.67	.51	.926
Reading	IV	Chile,Neth,Rum, Swe,USA	11.51	4.46	1.34	.666
		Eng,NZ,Scot,USA	9.60	5.86	.77	.935
Listening	IV	Chile,Rum,Swe, USA	12.83	3.36	1.47	.515
		Eng,NZ,Scot,USA	11.98	4.27	.85	.859

*Data from Sweden Population III, which took the Population II tests, were used in these comparisons.

such deviation in an across-English-speaking-countries analysis (-2.32 for England, Population IV Listening Item 38, a difficult item in any case).

Population II Reading:

Item 1: Rumania, +2.42.[7]
Item 9: Netherlands, +2.01.
Item 16: USA, +2.19.
Item 20: Rumania, -2.00.
Item 26: Rumania, -2.14.

Population II Listening:

Item 1: Rumania, -2.11.
Item 3: Rumania, -2.36.
Item 6: Rumania, -3.21; Sweden, +2.40.
Item 31: Rumania, +2.33; Sweden, -2.08.

Population IV Reading:

Item 1: Chile, -2.76; Rumania, +3.27.[7]
Item 5: Chile, -3.55.
Item 6: Chile, +3.77; Netherlands, -2.13; USA, -3.44.
Item 8: Chile, +3.32.
Item 10: Netherlands, +2.50; USA, -2.90.
Item 14: Rumania, +2.74.
Item 15: Chile, +3.12; Sweden, -2.14.
Item 16: Chile, +2.39.
Item 17: Chile, +3.38.
Item 28: Rumania, -2.57.

Population IV Listening:

Item 1: Chile, -2.14; Rumania, +2.98; USA, -2.13.
Item 2: Chile, -2.48.
Item 3: Chile, -2.55; Rumania, +2.26.
Item 4: Chile, +5.64; Netherlands, -4.50.
Item 17: Chile, +3.45; Netherlands, -4.62.
Item 26: Chile, +2.70.
Item 29: Netherlands, +2.84.
Item 37: Chile, +2.22; Rumania, -2.10.
Item 40: Netherlands, +2.13.

Explanations of these deviations would be only speculative; they could arise, for example, from characteristics of instructional materials in particular countries, as well as from linguistic differences. A suggestion that linguistic differences can account for the results comes from consideration of some Population IV Listening items in which the Chilean sample did better than expected. Item 4 has the script: "Le professeur va à la porte"--almost the exact counterpart of the Spanish "El profesor va a la puerta." The scripts for Items 17 and 26 also happen to contain forms of the French verb *aller (vais, va)* which closely correspond to Spanish forms. Further study of these matters would require a bit of contrastive linguistic research, but it would appear that most of the deviant results arise from problems with cognate words. On the whole, the tests seem equally appropriate for the various mother tongues.[8]

Anchor items. There were 15 items that were common to the Population II and Population IV Reading tests, and 19 items common to the Population II and Population IV Listening tests (although they did not necessarily come in the same order in the two tests). To see whether these items functioned similarly in the two populations, between-population correlations of the mean deltas for these items (averaged over the four English-speaking countries) were obtained. Also, the deltas of these items in the respective populations were compared with the deltas of the remaining items in the respective tests. The between-population correlations of the deltas were .76 for Reading, and .63 for Listening--not as high as might be expected, but clearly significant, statistically. Following is information on the comparison of deltas of anchor test items with those of other items in the respective tests:

	Reading				*Listening*			
	Pop. II		Pop. IV		Pop. II		Pop. IV	
	Mean	Sigma	Mean	Sigma	Mean	Sigma	Mean	Sigma
Anchor items	13.56	1.42	6.78	1.33	14.86	1.51	10.87	1.67
Other items	13.16	1.88	11.37	2.13	13.86	1.93	12.98	3.13
t	+0.67	n.s.	-4.74	($p<.001$)	+1.72	n.s.	-3.13	($p<.01$)

Thus, the anchor items are representative of other items thought to be appropriate to Population II and are quite easy for Population IV. As intended, the other items included in Population IV tests are considerably harder.

Order of Difficulty. The item analysis results show that very roughly the items tend to get more difficult in the latter portions of the tests. This rough correlation was as intended; it was not considered desirable to place the items in strict order of difficulty.

Scaling

The presence of a separately-scored set of "anchor items" in the Population II and the Population IV tests (both in the Reading tests and the Listening tests) made it possible to perform a vertical equating of the tests, that is, to estimate a way of assessing the results on a common scale. For this purpose, use was made of a technique developed by Levine (1955) for unequally reliable tests where the equating form U is included in both tests to be equated, and where the equating groups may be of unequal ability and are presumably random samples from their populations (Angoff, 1971, p. 582). Various preliminary analyses of the data suggested that the equating should be done by pooling information from New Zealand, Scotland, and the United States, these countries having samples from both Populations II and IV that seemed adequately representative, tested on both Reading and Listening. (Chile and Sweden were not tested at Population II, and the Netherlands was not tested in Listening. The sample from Rumania Population II seemed quite unrepresentative.) The data for the equating process and the results are shown in Table 11. The formulas for the final lines of equivalence, applicable to any score on the Population II test, are as follows:

Reading (using formula-correlated scores)

$$X' = 1.0908\ X_2 + 10.9433$$

Listening (using formula-corrected scores)

$$X' = .9742\ X_2 + 12.9397$$

These formulas contain a built-in constant of 20 such that negative results will be avoided. Thus, the score on the Population IV test is incremented by 20 in order to place it on the common scale, X'. Table 11 also shows the means of Reading and Listening tests for all countries, both populations, on this common scale. One may note from this the relative competence of the samples from the two populations. (England Population IV seems to have made greater progress than those in the other

Table 11

Equating Population II and IV Reading and Listening Tests (Formula-Correlated Scores)

	Reading		Listening	
	Pop. II	Pop. IV	Pop. II	Pop. IV
I *Basic Data* (From pooling of data from New Zealand, Scotland, USA)				
N (not used in computation)	7077	4585	4551	2703
($\bar{X}_p$) Mean Score, Total Test	9.411	19.981	5.984	14.912
(σ_p) SD, Total Test	8.351	8.202	7.252	10.390
($\bar{X}_{ep}$) Mean Score, Equating Items	3.529	11.353	1.544	9.379
(σ_{ep}) SD, Equating Items	4.163	3.975	3.901	5.405
(r_p) Correlations of Total Test & Equating Items	.912	.860	.879	.933
II *Parameters of Line of Equivalence*				
A (slope)	1.0908		.9742	
B (intercept; includes an added 20 points to avoid negative scores)	10.9433		12.9397	
III *Means of all countries on the common scale* $X'=A\bar{X}_2+B$, *or* $X'=\bar{X}_4+20$				
Chile	--	--	--	(23.20)
England	21.14	52.13	22.31	47.92
Netherlands	24.92	46.27	--	--
New Zealand	24.20	47.36	17.33	32.79
Rumania	40.21	48.62	36.38	46.24
Scotland	25.26	45.23	21.52	37.28
Sweden	--	39.60	--	39.30
USA	19.21	37.53	18.91	33.79

IV *Levine's equating formula (Case IV)* as modified for the above

$A=B_{e2}/b_{e4}$, where $b_{ep}=r_p\sigma_{ep}/\sigma_p$; $p=2, 4$

$$B=\bar{X}_4+\frac{(\bar{X}_{e2}-\bar{X}_{e4})}{b_{e4}}-A\bar{X}_2+20$$

countries; it should be remarked, however, that the England Population IV was an older, more advanced group in terms of the number of years of French instruction.)

Intercorrelations

An analysis of the intercorrelations among the several tests of French language competence can indicate the extent to which they measure, on the one hand, a general factor of proficiency and, on the other, separate aspects of proficiency. The correlational data available have limitations, however, in that in no country were the tests given to all or nearly all students in the samples tested. In some countries, no correlational data at all were available for certain pairs of tests, and in others, they were available for such small numbers of students that correlational analysis was deemed not worthwhile. Thus, no correlational analysis was attempted for data from the Netherlands, Rumania, or the United States. The matrices of correlations from the remaining countries still suffer from severe "missing data" problems, most generally with respect to the measures of Speaking Fluency, which were available for only 5 to 40 percent of the cases. Available for all these countries and populations are matrices for seven variables, i.e., Reading, Listening, Writing I, Writing II (Quantity and Quality scores), and Speaking Fluency (Quantity and Quality scores), on "weighted data," i.e., with cases weighted according to a probability sampling stratification scheme. In addition, matrices are available for New Zealand and Scotland (both Populations II and IV in each case) for the variables just mentioned and three additional variables, namely, Pronunciation, Structural Control, and Oral Reading, but for "unweighted data," i.e., for cases directly as they came. The correlations are in every case computed for all available pairs of scores for any given pair of variables; thus, some correlations are based on a larger number of cases than others.

Space does not permit presenting all the correlation matrices; the "unweighted data" matrices from Scotland, however, are shown in Table 12. These matrices were selected for presentation because they suffer least from missing data problems.

All correlation matrices that were available were subjected to simple factor-analytic procedures. (It was believed that any more sophisticated analysis would not be worthwhile in view of the limitations of the data.) First, principal-axis analysis was applied (with unities in the diagonals) to obtain n orthogonal components of total variance, where n is the number of variables. Second, the first four components were rotated by Kaiser's (1958) normalized varimax procedure. The decision to rotate four factors was somewhat arbitrary, but was influenced by the fact that the fourth principal axis was generally the last one to contain sizable loadings on more than one variable; in any case, it was thought advisable to treat all matrices uniformly.

The results of applying these procedures to the matrices of Table 12 are shown in Table 13, for the purposes of illustration.[9]

Loadings of the variables on the first principal axis for each matrix, shown in Table 14 along with the associated eigenvalues or roots, indicate the extent to which each variable tends to measure the same dimension as the total array of variables. From the medians of these loadings, we see that Writing I and Reading tend to have the highest loadings, followed by Listening, Writing II Quantity, Speaking Fluency Quantity, Structural Control, Writing II Quality, Speaking Fluency Quality, Oral Reading, and Pronunciation, in that order. All the tests, particularly those with the highest loadings, tend to measure a common factor of French language competence. It may be noted that this first principal axis accounts for anywhere from 46 to 69 percent of the total variance. It may also be of significance that for the three countries in which both Population II and Population IV data were analyzed, the percentage of total variance accounted for by the first principal axis was somewhat larger in the case of Population II, a finding that suggests that in Population II, the skills measured were less differentiated than in Population IV.

Examination of the four-factor varimax rotations suggests that to the extent that skills are differentiated, there are four fairly consistent factors across countries and populations:

1. A general language competence factor that is most prominently measured by

Table 12

Correlations Among French Proficiency Tests, Scotland
("Unweighted Data")

Population II
(*N*=about 800 except for variables 6-10, where it is about 200)

		1	2	3	4	5	6	7	8	9	10
Reading	1										
Listening	2	.76									
Writing I	3	.87	.72								
Writing II Quant.	4	.71	.64	.75							
Writing II Qual.	5	.68	.61	.71	.85						
Speaking Flu. Quant.	6	.66	.67	.65	.60	.49					
Speaking Flu. Qual.	7	.49	.43	.49	.49	.42	.70				
Pronunciation	8	.30	.38	.25	.21	.15	.28	.32			
Structural Control	9	.68	.71	.67	.54	.48	.66	.42	.23		
Oral Reading	10	.66	.62	.60	.53	.45	.60	.38	.35	.54	

Population IV
(*N*=about 975 except for variables 6-10, where it is about 235)

		1	2	3	4	5	6	7	8	9	10
Reading	1										
Listening	2	.68									
Writing I	3	.78	.64								
Writing II Quant.	4	.52	.46	.59							
Writing II Qual.	5	.53	.43	.60	.77						
Speaking Flu. Quant.	6	.56	.53	.53	.47	.40					
Speaking Flu. Qual.	7	.62	.54	.58	.48	.49	.69				
Pronunciation	8	.27	.36	.29	.27	.22	.33	.29			
Structural Control	9	.64	.64	.69	.53	.50	.59	.58	.39		
Oral Reading	10	.39	.44	.42	.37	.30	.42	.39	.31	.41	

Reading and Writing I, but also by other tests more or less in the order suggested above. This factor represents the extent to which the student has a broad and at the same time an exact knowledge of the grammar and vocabulary of the language, particularly in its written form.

2. A writing fluency factor that is measured by both Writing II Quantity and Writing II Quality (usually, better by the Quality score than the Quantity score), but very little by the other variables. This factor represents the extent to which the student can mobilize his language competence to produce a grammatically and lexically acceptable *written* composition. The extent to which this factor may be specific to the particular stimuli used in the Writing Fluency test cannot be determined from the data.

3. A speaking fluency factor that is measured both by Speaking Fluency Quantity and Speaking Fluency Quality (usually, better by the Quality score than the Quantity score), but very little by the other variables (with the possible exception of the Structural Control score). This factor represents the extent to which the student can mobilize his language competence to produce a grammatically and lexically

Table 13

Factor Analyses of Scotland Correlation Matrices

Population II

Principal Axis Factor Matrix

	I	II	III	IV	V	VI	VII	VIII	IX	X
1	.90	.09	.14	-.04	-.02	-.31	.07	.02	-.07	.24
2	.86	-.06	.21	-.11	.18	.16	.33	.16	.09	-.03
3	.89	.18	.08	.00	.04	-.32	.00	-.10	.05	-.23
4	.84	.31	-.03	.31	-.03	.15	-.09	-.15	.22	.09
5	.78	.41	.01	.38	.02	.14	-.04	.11	-.22	-.05
6	.82	-.16	-.34	-.22	-.04	.14	.14	-.27	-.12	.00
7	.65	-.27	-.66	.11	-.03	-.09	-.04	.20	.06	.00
8	.39	-.77	.27	.39	.13	.00	-.06	-.06	-.03	.00
9	.78	-.02	.08	-.43	.31	.09	-.30	.05	-.01	.01
10	.74	-.17	.24	-.18	-.56	.08	-.08	.08	.01	-.03
	6.05	1.03	.76	.68	.46	.31	.25	.19	.13	.12

Varimax Rotation of First Four Factors

		A	B	C	D	h^2 (Communality)
Reading	1	.67	.57	.20	.14	.84
Listening	2	.73	.43	.16	.23	.80
Writing I	3	.62	.63	.22	.07	.83
Writing II Quant.	4	.33	.84	.26	.07	.89
Writing II Qual.	5	.25	.91	.17	.03	.91
Speaking Flu. Quant.	6	.60	.24	.67	.07	.87
Speaking Flu. Qual.	7	.17	.25	.91	.17	.95
Pronunciation	8	.17	.05	.14	.96	.97
Structural Control	9	.83	.21	.24	-.01	.80
Oral Reading	10	.72	.26	.13	.27	.67
Variance		3.17	2.67	1.58	1.12	8.52 (Total Variance)

Factor A: General French Language Competence
Factor B: Writing Fluency
Factor C: Speaking Fluency
Factor D: Pronunciation

Table 13 *(continued)*

Population IV

Principal Axis Factor Matrix

	I	II	III	IV	V	VI	VII	VIII	IX	X
1	.83	.07	.26	.13	-.21	-.25	-.02	.18	.15	-.24
2	.78	-.16	.15	.10	-.30	.15	.45	-.07	-.05	.05
3	.85	.14	.12	.09	-.25	-.15	-.22	.10	-.13	.28
4	.75	.41	-.37	-.10	.09	.12	.09	.11	.27	.12
5	.72	.51	-.34	-.02	.06	-.05	.04	-.10	-.26	-.16
6	.75	-.19	.22	-.05	.46	.17	.02	.30	-.13	-.02
7	.78	-.04	.24	-.02	.41	-.23	.04	-.32	.08	.07
8	.46	-.59	-.56	.32	.05	-.14	-.01	.03	.00	.00
9	.82	-.09	.07	.17	-.10	.37	-.30	-.19	.05	-.08
10	.58	-.34	-.11	-.70	-.17	-.04	-.06	-.03	-.01	-.03
	5.50	.99	.79	.67	.63	.36	.35	.30	.21	.19

Varimax Rotation of First Four Factors

		A	B	C	D	h^2 (Communality)
Reading	1	.82	.31	.04	.06	.78
Listening	2	.75	.19	.23	.17	.68
Writing I	3	.74	.45	.08	.09	.76
Writing II Quant.	4	.31	.85	.11	.19	.87
Writing II Qual.	5	.32	.88	.06	.06	.89
Speaking Flu. Quant.	6	.73	.13	.13	.30	.65
Speaking Flu. Qual.	7	.75	.23	.05	.21	.66
Pronunciation	8	.20	.10	.96	.12	.98
Structural Control	9	.74	.30	.27	.10	.72
Oral Reading	10	.28	.16	.13	.92	.96
Variance		3.73	2.03	1.10	1.08	7.94 (Total Variance)

Factor A: General French Language Competence
Factor B: Writing Fluency
Factor C: Pronunciation
Factor D: Oral Reading (Note: A Speaking Fluency factor would have emerged if five factors had been rotated.)

Table 14

First Principal Axis Loadings, by Country and Population

Country	Chile	England	England	New Zealand	New Zealand	New Zealand	New Zealand	Scotland	Scotland	Scotland	Scotland	Sweden	Median
Population	IV	II	IV	II	II	IV	IV	II	II	IV	IV	IV	
Reading	.65	.87	.77	.84	.81	.65	.60	.89	.90	.85	.83	.85	.835
Listening	.72	.82	.79	.66	.71	.73	.78	.83	.86	.77	.78	.83	.783
Writing I	.82	.85	.80	.86	.80	.81	.72	.90	.89	.86	.85	.85	.859
Writing II Quant.	.81	.84	.57	.81	.74	.71	.60	.87	.83	.77	.75	.78	.777
Writing II Qual.	.46	.82	.67	.80	.71	.71	.63	.83	.78	.76	.72	.68	.716
Speaking Flu. Quant.	.82	.78	.59	.70	.72	.68	.75	.82	.82	.74	.75	.73	.747
Speaking Flu. Qual.	.73	.69	.49	.66	.68	.70	.71	.67	.65	.79	.78	.73	.698
Pronunciation	--	--	--	--	.41	--	.52	--	.39	--	.46	--	.434
Structural Control	--	--	--	--	.67	--	.65	--	.78	--	.82	--	.724
Oral Reading	--	--	--	--	.62	--	.66	--	.74	--	.58	--	.641
Root	3.68	4.61	3.22	4.10	4.84	3.58	4.41	4.84	6.05	4.41	5.50	4.29	
Percent of trace	52.5	65.9	46.0	58.6	48.4	51.1	44.1	69.2	60.5	62.9	55.0	61.3	

adequate *spoken* response to the test stimuli. As in the case of the writing fluency factor, the extent to which the factor is specific to the particular stimulus situation presented in the Speaking Fluency test cannot be determined from the data. (It might represent, for example, the student's reaction to the requirement of recording his response on tape, and not have much to do with his ability to carry on a conversation in French.)

4. A spoken language competence factor measured mainly by the Pronunciation test and the Listening test, also to some extent by the Oral Reading test. The factor seems to represent the extent to which the student's competence in the phonology of French enables him to perform well on tests requiring him to produce French with accurate pronunciation, and also helps in the comprehension of spoken French. Because correlations for the Pronunciation and Oral Reading tests were available only for two countries, the nature of this factor is not clear. The analyses for the seven-variable matrices (that included Listening but not Pronunciation or Oral Reading) do suggest, however, that listening is a clearly differentiated skill that stands somewhat apart from reading, writing, and speaking.

SUMMARY

This chapter has presented the considerations that went into the design of the French proficiency tests used in the IEA study, descriptions of the tests themselves and of the manner in which they were administered and scored, and analyses of their technical characteristics as revealed in the voluminous data that were collected in the course of the study. The analyses show, it is believed, that the tests were fully adequate for the purposes for which they were designed. They have adequate reliabilities, and examination of their intercorrelations shows that they agree in measuring a general language competence factor, as well as several differentiated skills. They appear to be highly, and about equally, valid for measuring French proficiency in all the countries that participated in the study.

FOOTNOTES

[1] The International Association for the Evaluation of Achievement (IEA) is an organization of representatives from twenty-one countries who joined together in the 1960s to undertake international studies of educational achievement in different subjects. French as a Foreign Language is only one of the studies undertaken. Other subjects were: English as a Foreign Language, Science, Reading Comprehension, Literature, and Civic Education. Headquarters of IEA are at the Institute for the Study of International Problems in Education, University of Stockholm, Fack, S-106 91 Stockholm 50, Sweden. Application to this institute may be made for permission to use or adapt the French proficiency tests described in this chapter. The tests themselves are available in microfiche or photocopy as Document 102-184 from the ERIC Document Reproduction Service, P.O. Box 190, Arlington, Virginia, 22210.

[2] A complete report of the French as a Foreign Language study has been published (Carroll, 1975).

[3] Despite complaints from some countries that the rate of speech in pilot testing materials was too rapid, an effort was made to voice the French in the natural, idiomatic manner typical of normal conversations or broadcasts. In giving illustrations here, we indicate the actual time, in seconds, from the beginning to the end of the voicing.

[4] One other variable had been used in the scoring of pilot testing responses, namely, number of clauses exhibiting one or more hesitations, but this seemed not to be of any special value in predicting global ratings; therefore, it was not included in the final scoring procedures.

[5] See fn. 2.

[6] The global Fluency ratings were used only in determining weights for components of the total Speaking Fluency scores; they were not used as criterion variables in final data analyses.

[7] Item 1, Population II Reading test happens to be an "anchor" item that is the same as Item 1, Population IV Reading. Rumanian students excelled in both instances,

possibly because the item involved the French verb *manger*, "eat," corresponding to *mînca* in Rumanian. The Chilean students tested at Population IV may have excelled because they recognized *manger* from the Spanish word, *manjar* ("food"). However, other "anchor" items in the above list did not show similarly consistent results.

[8] Unless one believes that the influence of linguistic similarities pervades the material in the test as a whole. It could be argued, that is, that perhaps Rumanian students did better than those in other countries because Rumanian is closer to French than the other languages are. But Spanish is probably even closer to French than Rumanian, and the Chilean sample did not do at all as well as those in English-speaking countries, and far less well than the Rumanian sample. Our general results suggest that the poor overall performance of Chilean students was due to nonlinguistic, educational factors--less instructional time devoted to French, less motivation to study French, and so on.

[9] Nevertheless, these matrices are somewhat atypical in their factorial structure. The factor interpretations presented below are based on a consideration of the complete array of results.

IEA International Study of Educational Achievement Test of Proficiency in English as a Foreign Language

E. Glyn Lewis, Wales

ORIGINS AND PROGRESS OF THE INVESTIGATION

In 1965 the International Association for the Evaluation of Educational Achievement (IEA) inaugurated a cross-national study of achievement in 6 subjects--Science, Reading Comprehension, Literature, Civics Education, French as a Foreign Language, and English as a Foreign Language. It had embarked upon a study of Mathematics (Husen, 1967) some time earlier, and the results of the study were known before the Six Studies programme had begun. Over 20 countries have been engaged in the total IEA project, though all of them have not participated in every study. The countries participating in the English study are Belgium (French Region), Chile (Population IV only), Federal Republic of Germany (FRG), Finland, Hungary (Population IV only), Israel, Italy, the Netherlands, Sweden, and Thailand. The general aim of the IEA studies has been to arrive at useful conclusions about the ways in which selected factors affect educational achievement and to estimate the relative functional value of differing educational systems.

The fact that the achievement tests in English were developed with a view to allow conclusions to be drawn which were of a more general nature than affect English specifically, and which are to that extent subordinate to a very general plan of investigation, does not entail that the criteria by which the tests should be judged are less rigorous than would be the case if we were concerned with English alone. For this reason an International Committee, drawing upon the resources of 6 countries, spent over 4 years in designing and developing appropriate cognitive and other measures. The writer has served as Chairman, the other members being Ian Dunlop (Sweden), Esther Heitner (Israel), E. Morette[1] (France), Agnes Niyakawa-Howard (United States), Rauno Piertola (Finland), and Clifford H. Prator (United States). The International English Committee was supported by, and worked in close co-operation with, English Committees in each of the participating countries. The National Committees produced a contents analysis of the English courses in their respective countries, indicated the English teaching objectives of the courses, and estimated their relative importance. They also supplied samples of items, many of which, after being modified, were included in the final tests. The investigation was administered by the IEA organisation whose technical officers, especially the Executive Director, Neville Postlethwaite, and his successor, Roy Phillips, guided the International Committee and undertook the very specialised socio-metric aspects of the study.[2]

The International Committee met for the first time in November 1966 and considered the hypotheses to be tested and the variables about which information would be required. The Committee decided on the general specifications for the tests and set out a pro-forma which the National Centres were invited to use in indicating their teaching objectives and their relative importance (see Tables 1 and 2). In October 1967 the International Committee met again to consider the observations of the National Centres. In the meantime some members of the Committee prepared a.

provisional set of tests and questionnaires. These were considered by the Committee, and their conclusions were referred again to the National Centres. The French and English Committees co-operated closely in these developments. In November 1968 the International Committee met for the third time to consider the results of the pretesting of the provisional instruments, which had occurred during the spring of that year. The tests were again modified in the light of the pilot testing score distributions and item analysis data which had been assembled by the Technical Officers. In October 1969 the Committee met for the fourth time to review the results of the second pretesting of instruments revised during the third meeting and during sessions in which 1 or 2 members of the Committee met the Technical Officers. Each pretesting programme involved approximately 2,000 students drawn from several countries. The 1969 revised tests and other measures were submitted to the National Centres and were then approved by the Council of IEA at the beginning of 1970. Most of the tests as they were published consisted of some original items, some supplied by National Centres and modified by members of the Committee, and some obtained from released sources in Research Organisations and Institutes. In some of the multiple choice items, the behaviour of the offered alternative answers varied from country to country. Where this difference did not appear to affect the validity of the item, it was not considered to be a serious disadvantage, and probably inevitable in any case. The tests for the 2 "target populations," it was decided, should conform to the same basic pattern.

The tests, questionnaires, and manuals of guidance on the administration and scoring of the tests, where this was not undertaken centrally, were published and distributed to the National Centres in 1971. Where it was necessary, the National Centres translated into the native language test item instructions to students, as well as instructions to administrators. They processed the taped recordings necessary for the Listening tests and organised the conduct of the investigation in their own schools. The period between 1971 and 1973 was devoted to scoring the tests and collecting the data produced by other measures, undertaken partly at Teachers College, Columbia University, and partly at the University of Stockholm. The Speaking and Writing tests were scored and reported upon at the National Foundation for Educational Research (London), under the direction of Dr. Clare Burstall.

THE SIGNIFICANCE OF THE IEA FOR THE STUDY OF ENGLISH LANGUAGE PEDAGOGY

Since the English tests were a necessary but subordinate element in the overall investigation, and since the achievement tests were only a part of the English programme, a true picture of the English tests and a fair analysis of their aims can only be derived from an account of how the other measures--school, student, and teacher questionnaires, as well as studies of the economic and other characteristics of the participating countries--were developed and used. Since this would require a much lengthier discussion than is appropriate here, the limitation will have to be accepted for all its inconvenience. However, there are some considerations which need to be discussed at this stage.

Because the cognitive measures were a subordinate element in the English investigation, their constitution was determined as much by the overall design of the total IEA programme, which was oriented towards an investigation of sociological factors in the main, as by the requirements of language testing *per se*. This did not necessarily embarrass the members of the English Committee. Some of them held the view that the most important aim of such testing was to improve language pedagogy--including administrative arrangements for the subject, the training of teachers, and the preparation of materials. It was felt that the investigation of English in the context of IEA philosophy and IEA research might contribute to the understanding of the sociological factors involved in teaching English, and that this might lead to considerable rethinking of the nature of achievement in English and a reconsideration of what tests are necessary to measure that achievement. Hitherto, the changes which have been introduced into second and foreign language pedagogy have been determined less by advances in understanding the operation of social factors than by developments in psychology and linguistics. The influence of the home and the neighbourhood, together with a country's social and economic aims, have usually figured fairly prominently in discussions concerning the curriculum as a whole (the macro-structure) but not so prominently in thinking about the teaching of specific subjects (the micro-structures) and even less in the case of foreign language teaching than in the case of teaching the mother tongue and the sciences. At the same time the aims of teaching foreign languages have changed very considerably: they are governed less by the interests of academic discipline (traditional philology), or even humanistic and broadly cultural values, and increasingly by more practical objectives relating to the needs of immediate, face-to-face communication. This tendency has been reinforced by the movement to make foreign languages available not only to selected groups or an elite but to the majority of students in all schools and in some cases to all students from an early age.

TARGET POPULATIONS AND SAMPLING

The English Committee, conforming to the wishes of the IEA overall planners, was originally interested in 3 age groups--10-year-old, 14-year-old, and pre-university grade students. The intention to test the first group was abandoned because a sufficiently representative sample from enough countries could not be obtained; that is, large enough samples of 10-year-old students who had pursued a course of English for at least 18 months could not be found in the participating countries, and it would have been uneconomical, if not impossible, to test all the students in the remaining groups. In any case it was not necessary to do so. The success of the present study depends on the samples chosen being a good estimate of the total population. Because most countries have a policy of promoting students according to age, it was necessary to draw samples from specified age groups. However, the use of the age sample posed some problems for those countries where promotion was determined by achievement, because students of the same age might be found in different grades. It was also necessary to test and question in the pre-university group those who specialised in the subject, as well as those who had studied it for at least two years at any point in their school careers but were not at present pursuing English courses, or if they were, did not propose to specialise. The definition of the Target Populations is the same for all subjects.

Population II consisted of all students of English aged 14.0 to 14.11 in full-time schooling but excluding certain categories such as mentally handicapped or disturbed children. This Population was decided upon because it was the last point in the school systems of most countries where 100% of the age group remained in school. The second target group, *Population IV*, consisted of all students in the pre-university grades in full-time schooling who were either specialising in English (*IVs*) or had previously studied it for at least 2 years. In addition to these 2 groups of students, 4 countries included an optional group, *Population III*, which consisted of students in full-time schooling at a major terminal point in the school system beyond the end of compulsory schooling and before a university or comparable education begins. The composition of this group differed from country to country. The cognitive tests administered to Population III were identical with those for Population II, although supplementary national options were included in some countries.

The sampling procedures for English were complicated by the fact that in most countries English is an optional subject. Only some schools offer the subject, and even those that do offer it to only a selection of students. There was a lack of initial information about which schools contained students taking English as a foreign language. Consequently, the first step was to draw stratified samples of schools that were large enough to leave a sample of reasonable size after the selected schools subsequently found not to contain such students had been discarded. If the sampling design kept a constant overall sampling fraction, every student taking English would have the same chance of appearing in the sample. Each National Centre prepared its own sampling plan, which was checked by the international sampling referee (G. F. Peaker) before it was finalised and the schools approached. Although each sample was sent to the IEA sampling referee for approval, the responsibility for the quality of the sample and the data in the main analysis rested with each National Centre.

TEST SPECIFICATION

Two questions concerning the nature of the proposed tests were considered. First, how far would they seek to encompass the different aspects of the language--phonology, grammar, vocabulary, etc.? Second, would they be selected so as to assess competence in all the skills--Listening, Reading, Speaking, and Writing--or only some of them, and if so, which ones? These 2 considerations became the co-ordinates of a frame within which the tests were constructed and concerning which the National Centres were invited to make their observations. The following pro-forma was circulated to them:

Table 1

Specification

Level of Integration	Kind of Channel	Decoding	Encoding
Level 1	Auditory-vocal	Identification and discrimination of sounds	Pronunciation, intonation, and stress
	Visual-kinetic	Identification and discrimination of letters	Spelling, punctuation, and capitalization
Level 2	Auditory-vocal	Recognition of letters, words, and sentence patterns	Active vocabulary; production of grammatical units
	Visual kinetic		
Level 3	Auditory-vocal	Auditory comprehension	Speaking
	Visual-kinetic	Reading comprehension	Writing

Phonology

The Listening and Speaking tests would assess the student's acquaintance with the sound system of English, including segmental units such as vowels and consonants, together with their relationships and distributional pattern, as well as intonation. Phonological elements would be tested, not in isolation but as parts of meaningful stimuli. In accordance with the views expressed by all the countries, the model to be used would be British rather than American English.

Writing

Students would be tested for their ability to understand the Roman script and English orthography. In addition to Reading tests, exercises in dictation would be set. The understanding of different kinds of prose, complex and abstract styles, as well as the ability to produce free composition, would be tested.

Grammar

Students would be tested for their ability to understand and use grammatical signalling devices, such as inflectional affixes, word order, and functional operatives in simple and complex contexts. No specified selection or *a priori* system for grading structures was contemplated.

Vocabulary

The range of the student's vocabulary, both active and passive, would be tested in context. For the same reason that no specified system for selecting and grading structures was contemplated, the selection of the vocabulary would be undertaken in consultation with the National Centres.

Literature and Culture

The desirability of testing knowledge about the English language, English literature, and its related culture and history would be determined by the consensus of the participant countries. However, whether knowledge of English literature were tested or not, it was proposed to obtain information about attitudes to English, as well as student aims for learning the language, etc., by means of questionnaires.

To ascertain what the National Centres thought of these specifications, as well as to obtain information about the relative importance of possible objectives for English teaching, another pro-forma (Table 2) was issued. In the table, Column (1) indicates the objectives; Columns (2) and (4) rank order those objectives from 1 to 7 (major objectives) and from a-f (subordinate objectives); and Columns (3) and (5) enable the National Centres to indicate how important, on a 3-point declining scale, each of these major and subordinate objectives are for Populations II and IV separately. The table indicates the response of one country only.

The areas where disagreement was most apparent among the countries is that of developing favourable attitudes. There was hardly any difference in the responses in respect to the 2 Populations separately, except that some countries gave greater prominence in Population IV than in Population II to reading passages of different levels of abstraction, and less prominence to reading short passages intensively. In the second place, some countries gave greater prominence in Population IV than in Population II to free composition, reproduction, and summarizing and the recording of oral conversation. All these changes in emphasis were taken to represent an expectation in the National Centres that the level of ability in English would be higher in Population IV than in Population II.

PROBLEMS OF TEST DESIGN AND DEVELOPMENT

Theoretical Problems

The considerable differences known to exist between the countries in respect of the place of English affected the development of the tests and questionnaires. For instance, it was envisaged initially that the English study would embrace such countries as the United States, Australia, and Britain, as well as the 10 already listed. In these 3 countries English is taught to a few students as a foreign language, to many more as a second language, and to the great majority as the mother tongue. Had these countries participated, an additional major factor would have affected the test design. For instance, it would have been possible, and perhaps necessary, to test children well below 14 years of age, and the range of test difficulty for the 2 existing target populations would have had to be extended well beyond its present limits.

The fact that, in the 10 countries which continued to participate, English is taught as a foreign rather than as a second language reduced, but did not eliminate, the complexity of the situation. In some countries English is used to teach some other subject(s), if only to a slight degree. Only some of the 10 countries share historical associations with the English speaking world, so that differences in motivation for acquiring English may be reflected in the types of courses the students pursue. Such differences are associated with the fact that in some cases English is the first foreign language offered to students, in some cases it is the second foreign language, while in 1 or 2 countries English is a required subject for all students. In some countries where English is taught to the majority of secondary students, there may be a second indigenous national language which they are expected to acquire. The age at which English is introduced and the amount of time allotted to the subject each week, as well as the type of training undertaken by

Table 2

Behavioural Objectives for English

(1)	(2)	(3)			(4)	(5)		
		Population II				*Population IV*		
Objective		1	2	3		1	2	3
Listening	2	X			2	X		
Understanding spoken English in general	a	X			a	X		
Understanding different varieties of English	c			X	c			X
Understanding different registers--formal and conversational	b			X	b		X	
Speaking	3	X			3	X		
Speaking English generally	a	X			a	X		
Speaking different varieties	c			X	c			X
Speaking different registers	b			X	b		X	
Reading	1	X			1	X		
Reading for information at various levels, such as news items or science	a	X			a	X		
Reading at different levels of abstraction	c		X		b	X		
Reading for pleasure--literature at various levels of formality	c			X	c		X	
Reading quickly	d			X	d		X	
Reading short extracts intensively	b	X			c	X		
Writing	4		X		4	X		
Simple sentences	a	X			g		X	
Informal writing--recording a conversation	f			X	d	X		
Formal, e.g., a business letter	d		X		e	X		
Reproducing and summarizing	e			X	b	X		
Free composition--description, narrative, and exposition	c		X		a	X		
Frequent practice in writing	g			X	f		X	
Writing to instill accuracy	b	X			c			X
Knowledge about English	6			X	6		X	
History of the language	d			X	d			X
English literature	a		X		a		X	
English history	b			X	b			X
Acquaintance with the English way of life	c			X	c		X	
Developing favourable attitudes	5			X	5		X	
To the English language	a		X		a	X		
To English-speaking peoples	b			X	b		X	
To English culture and institutions			X				X	
Status of English in the curriculum								
as an academic study				X			X	
as a practical study			X			X		

teachers of English, differ greatly from country to country. If the tests of achievement were to be even minimally appropriate in such varying contexts, they would need to be as simple as possible. They had to be true tests aimed at as wide a range of achievement as the varying circumstances of English pedagogy would necessitate and had to refer to only the basic competencies. But the tests could not be designed to take these factors into account, since it was their operation that the total study was concerned with.

One of the basic considerations in the mind of the English Committee was that the tests were to be international in nature--not simply in the sense of being available for testing in different countries but basically to enable comparisons to be made between different countries. Yet teaching English presents different problems (or the same problems in different dimensions) to students in different countries. The 10 countries represented 10 different native languages, and there is no questioning the fact that this dissimilarity in native languages will produce different effects on the learning of subsequent languages. For this reason some of the 10 countries pressed for the adoption of "national tests" which would reflect the nature and extent of native language influence, but this was not considered feasible. Even if separate tests could be used in the present study, there are several reasons why the case for their adoption should be overruled. In the first place, to adopt the principle of "contrasting pairs" is to place a premium on the contribution which the theory of contrastive analysis may make to teaching. Not only do the countries involved in the present investigation differ in their recognition of its usefulness, but there is no consensus concerning its value among equally competent theoretical and applied linguists. Furthermore, to select the tests even partially on the basis of any one principle would tend to deprive the error analysis of much of its value, since in that case the probability of certain errors would have been determined already to some extent. One of the products of the testing programme, it was hoped, was the accumulation of information across countries about the incidence of certain types of errors, and the influence of the native tongue is only one source of such errors.

A second problem not unrelated to the first was the possible effect of cultural differences upon the interpretation of and response to the items and the pictorial material accompanying them. It was decided to obviate this by relating the nonlinguistic content of the tests to the lowest common denominator of culture--a "basic culture content" common to all countries. The disadvantage of adopting this procedure is that the items may appear mechanical and austere.

National differences of attitude to the idea of achievement testing at all is also relevant in considering the achievement scores of students in different countries. A similar type of difficulty arises from the fact that different countries adopt different attitudes to different specific types of tests and are accustomed to use some of them less frequently than others. For instance, the use of objective and multiple choice items is not favoured in some countries; dictation and translation are more important in others.

The vocabulary and structures selected for the items posed a considerable problem. French is more favourably placed than English in having a largely centralised organisation for the promotion of the language overseas and a general consensus on the norms of French which are to be adopted in teaching and testing, as well as agreement on the selection and grading of vocabulary and structures, based on fundamental research. In the case of English, however, there are only several major organisations in several countries concerned with the promotion of English, there are at least 2 major variants of the language, British and American, with their norm differences, and there is no agreement on whether to impose even one of those sets of norms or to accept "local variants." Because of these considerations, there is little agreement among course designers and textbook writers on the selection and grading of the vocabulary and structures of English.[3]

Finally, the Committee considered how far to go in distinguishing types of items--for instance, those which were primarily vocabulary items from those which were mainly structure items and, on the other hand, items which were distinguished according to the feature of grammar they might be testing.

Eventually, the English tests were organised so as to distinguish first between aspects of English--for instance, tenses or number or voice--and levels of the

English language, namely vocabulary, grammar, and phonological features, and second between types of test techniques, namely tests which were intended and designed as far as possible to identify a specific feature--the specific features of grammar, for instance--or to assess a more general acquaintance with the English language--for instance, the comprehension of sentences or longer passages where the success of the students depended on their understanding of several aspects of English equally.

The several groups of items exemplifying these differences or dichotomies in test organisation cannot be mutually exclusive. Although items which were meant specifically to assess control of identifiable grammatical features are excluded from the group meant to test control of vocabulary, items in both these groups could be included in either the "global" or "point-specific" type test. It was recognised all along that the dichotomies established for these tests could not be absolute and categorical--all tests made use of vocabulary, obviously, but it was felt that the tests were designed in such a way that when vocabulary was being tested, the grammatical component of the item would be minimised, and that when the grammatical features were being tested, the vocabulary component would be made so simple as to ensure that it did not constitute a significant test variable in that item. Similarly in the case of the second dichotomy, it was recognised that an item which was directed specifically at the understanding of a tense form, for instance, would need to be placed in a sentence the whole of which would need to be understood. Therefore, although a point-specific test is to some small degree also a test of general understanding of English, there is a marked difference between such items and those in the reading comprehension items, for instance, where the whole point of the item is to assess the ability to understand the purport of what is said.

Practical Problems

The international nature of the testing programme (and particularly the diversity of languages represented by the 10 participant countries) and the very large number of schools and students involved, together with the fact that many schools would be testing more than one subject, made it necessary to produce tests that could be administered easily and would take up as little time as was consonant with the requirements of validity and reliability. These considerations were reinforced by the expectation which was borne out in fact that the processing of such a large amount of data would not be easy. So far as concerns the timing of the tests, all 4 sets--Reading, Listening, Speaking, and Writing--were divided into sections according to the different categories and types of items. As a result of the 2 pretesting exercises, these sections were allocated sufficient time to enable the great majority of students to complete the test. The use of multiple choice testing was necessitated by the requirements already discussed, although several participants were anxious to avoid the method, for several different reasons. Differences in experience in handling such tests no doubt contributed to some of the between country differences in scores. Unlike the effect of other variables, the effect of this could not be taken into account in the present programme, but it should be in the minds of all those who will be interested in interpreting the results.

The Listening and Speaking tests involved the use of tape recorders and, apart from the fact that some of the 10 countries could deploy a very few of these among the schools drawn into the sample, many teachers, even if they were accustomed to using them in teaching, were fearful that they might do less than justice to their students in adopting tapes for testing. It was more difficult to handle the Speaking test than the Listening test, and it was decided for these reasons among others that the Speaking test (like the Writing test) should be an optional part of the total testing programme. All the non-objective items were scored centrally.

Generally speaking older students (Population IV) might be expected to have studied English for at least 3 and possibly 4 years longer than Population II. This generalisation was falsified to a considerable degree. For instance, Hungary introduced English in the 9th grade, while all Swedish students began English in the 4th grade and all Thai students in the 5th. Furthermore, there were considerable differences within the same country in some instances, partly because a system of school options enabled some students to take a foreign language relatively late and partly because they had the option to take up the study of English as a second foreign language almost 6 years after the course in English as a first foreign language began. In one country, while 70% of the students began English in the third grade, 27% began the course in the 9th grade. In Chile, between 5-10% began in grade 4 or 5, the majority in grade 7, and a further group in grade 9. These considerations affected assumptions about the proficiency levels at which the tests should be directed and especially about the existence of fairly equal opportunities to have attained such proficiency levels. The tests were prepared on the assumption that all students tested had followed a course in English for a minimum of 2 years. But there could be no guarantee that all the students who were included in the sample fulfilled this requirement because the sample was drawn in the main on the basis of age. Nevertheless, it is the only assumption that can be made about the length of experience of students included in Populations II or IV.

It would have been uneconomical and otherwise impractical (in the context of this project) to have had one set of tests for both Populations. Because the sample of students specialising in English turned out to be very small when compared with the other groups tested and because Population III was tested only in 3 countries and as part of a programme of "national options," it was decided to have 2 sets of tests only--Population II/III and Population IV/IVs. At the same time, in view of differences within each Population and the overlap between the 2 Populations in respect of the length of student experience of English, the range of difficulty of the 2 sets of tests would have to be very considerable, corresponding in Population II to 8 years to 2, and in Population IV from 12 years to 2. Some students in Population II had 4 years more experience of English than some students in Population IV. The overlap that was required in the range of difficulty of the 2 sets of tests was not entirely disadvantageous, although the possibilities of establishing comparisons between levels of achievement of the 2 Populations are made very much more difficult by the fact that neither Population is uniform from the point of view of experience of the language.

CONSTITUTION OF THE TESTS

Test Inventory

An inventory of the tests for both Populations is given in Table 3.

Description

Listening Tests: Population II. The stimulus material was recorded on tapes. Initially these were prepared at the Language Centre of the University of Essex (England), but changes in their format made it advisable to have them prepared in the United States using British voices. Copies of the final versions were sent to the participating countries, who prepared copies for the schools. Instructions were translated and were set into the tapes at appropriate points. Each student made his response in an accompanying booklet which also contained other necessary material such as pictures.

Section I of the tests for both Populations consisted of picture items. The sketch of a simple object was printed, and the student was required to identify which of the 3 spoken words corresponded to the picture (see *Item 2*). The words

(Item 2)

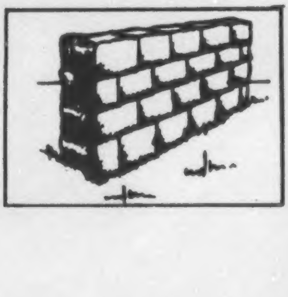


Table 3

	Population II				*Population IV*		
Type	Description of Items*	No.	Max. Score	Type	Description of Items*	No.	Max. Score
Listening				*Listening*			
Section 1	Discrimination of sounds	12	12	Section 1	Discrimination of sounds	12	12
Section 2	Comprehension of sentences	12	12	Section 2	Intonation	8	8
Section 3	Dictation			Section 3	Listening comprehension	8	8
				Section 4	Listening comprehension	8	8
Reading				*Reading*			
Section 1	Recognition of antonyms	6	6	Section 1	Recognition of word stress	8	8
Section 2	Recognition of pronunciation	8	8	Section 2	Collocations	8	8
Section 3	Grammatical structures	10	10	Section 3	Grammatical structures	8	8
Section 4	Definition of the meaning of words	6	6	Section 4	Comprehension--short sentences	11	11
Section 5	Reading comprehension--short sentences	14	14	Section 5	Comprehension--continuous text	25	25
Section 6	Reading comprehension--continuous texts	16	16				
Writing				*Writing*			
Section 1	Grammatical structures	13	13	Section 1	Grammatical structures--miscellaneous	13	13
Section 2	Grammatical structures			Section 2	Grammatical structures		
(a)	Pronouns	2	2	(a)	Pronouns	2	2
(b)	Plurals	3	3	(b)	Plurals	3	3
(c)	Tenses	3	3	(c)	Tenses	3	3
(d)	Voice	2	2	(d)	Voice	2	2
Section 3	Word Order	8	8	Section 3	Word Order	8	8
Section 4	Directed Composition	1	**	Section 4	Directed Composition	1	**
Speaking				*Speaking*			
Section 1	Grammatical structures with pictures	12	12	Section 1	Grammatical structures with pictures	12	12
Section 2	Oral Reading	1	30	Section 2	Oral Reading	1	30
Section 3	Fluency	1	**	Section 3	Fluency A	1	**
				Section 4	Fluency B	1	**

*Observations on the time allocation are made in the details of description.
**Explanations of scoring are included in the detailed description of items.

repeated on the tape were wall, ball, and fall. In some items the initial consonant was the significant element, as in the sequence illustrated on p. 57, sometimes it was the vowel--cut, coat, and caught--and sometimes it was the final consonant--cat, cap, and cab. All the words were monosyllables. The item analysis for both Populations in respect of the illustrated item is included in Table 4 as an example.

Table 4

*Item 2**

	Population II						*Population IV*					
Country	P-B	Omits	A	B	C	Nos.	P-B	Omits	A	B	C	Nos.
Belgium	.267	3.0	79.6	10.6	6.8	687	.294	1.8	88.7	6.3	3.2	1394
Chile							.376	7.9	67.3	17.2	6.3	231
FRG	0.0	0.0	100	6.0	0.0	6	0.0	0.0	0.0	100	0.0	1
Finland	.334	1.2	66.7	14.0	18.1	2118	1.76	0.5	91.4	5.0	3.1	2296
Hungary							0.19	0.1	86.4	6.5	6.9	1062
Israel	.112	34.4	3.3	46.8	15.4	17	0.58	35.3	28.2	32.5	4.1	10
Italy	.286	2.2	76.2	14.4	7.2	694	.384	1.0	84.7	11.4	3.0	314
Neth.	0.0	62.4	0.0	37.6	0.0	5	0.0	0.0	100	0.0	0.0	2
Sweden	.347	0.2	86.8	4.5	8.4	2268	.223	1.0	97.3	0.9	0.7	1601
Thai.	.534	45.2	22.6	4.8	17.8	12	0.0	0.0	100	0.0	0.0	1

*The item is identical for Population II and IV. A is the correct response. Chile and Hungary participated in testing Population IV only.

Item 2 was fairly easy for both Populations and easier for Population IV than for Population II. Both Populations tended to select the same incorrect response when they failed to choose A.

Section 2: Comprehension of Sentences. The students were required to listen to a sentence, or 2 very short sentences, and to select from the alternatives printed in their booklets the sentence which interpreted the stimulus sentence correctly. Two of the items consisted of short conversations. For instance, one such item was the following:

(Item 13)

The student heard:

John and Mary are at a table in a restaurant.
John says, "Where do you live now?" Mary says, "On State Street."
John says, "Why did you leave the High Street?"
Mary says, "I didn't like it." (Pause)

The booklet contained the following alternatives (correct response is indicated here by an asterisk):

A Mary lives in High Street.
B Mary lives in a restaurant.
*C Mary lives on State Street.**

Another item was a question arising out of a simple statement.

(Item 19)

The voice spoke:

Charles is having a party today to celebrate his sixth birthday.
How old is Charles? (Pause)

The booklet listed the following alternative answers:

*A Six years old.**
B Sick.
C Six boys and girls.

The percentage difficulty of these 2 items (*13* and *19*) varied in each case from 12% correct in one country to 92% in another, with the majority ranging around 56%.

Section 3 consisted of a Dictation, the passage consisting of 112 words. Each incorrect word was reckoned as an error. After the first 8 errors, each subsequent set of 8 errors meant the loss of 1 point from a possible maximum of 12.

Listening Tests: Population IV. The first section was identical with that for Population II, and the 12 items constituted part of the set of Anchor Items.

Section 2 was designed to test the ability to recognise difference in meaning arising from the different intonation patterns of identical sentences. The student was required to identify which sentences conveyed a certain type of meaning--for example, simple questions, questions requiring yes or no answers, incredulity, etc. Sometimes all 3 sentences, sometimes only 2 sentences, and at other times no 2 sentences conveyed the same meaning. Students did not have much difficulty in differentiating intonation patterns conveying simple questions, even though the grammatical structure was not strictly interrogative, e.g.:

(Item 13)

The taped sentences were as follows, with indicated intonation levels:

2 3
Father's taking a nap
2 3
Father's taking a nap
2 3
Father's taking a nap

The students were asked which of the sentences were questions, and the answer was that all 3 were.

Students found it more difficult to distinguish sentence intonation patterns which were slightly more complex, such as those concerned with questions referring to time as in the following:

(Item 14)

The taped sentences were:

2 3
Did she work in a shop before Christmas
2 3
*Did she work in a shop before Christmas**
2 3
Did she work in a shop before Christmas

In Section 3 the student was tested for his ability not simply to understand what was said but to draw correct inferences. Short sentences were used in some instances like:

(Item 21)

The taped sentence was:

If we had known we could have ridden we would have gone.

and the alternative responses printed in the booklet were:

A We went.
*B We did not go.**
C We knew that we could go.

Most of the 8 items in this section proved to be difficult except in those countries where the general level of achievement in English was fairly high. Many of the items consisted of 3 or 4 sentences like the following:

(Item 27)

The students heard:

"In English schools music is one of the principal studies. The schools are trying to encourage musically talented youngsters. They also give those who lack such ability a chance to acquire a taste for fine music through courses in music appreciation."

The listed alternative responses were:

A Music education is given only to the talented.
*B Music education is given to the talented and non-talented.**
C The schools do not encourage talented children.

The objection had been raised that some of these items test the student's memory as much as they do his ability in English. However, success in these items does not depend on the ability to retain detail so much as the ability to come to a conclusion which is conveyed by the general sense of the passage. This type of item (*Items 26, 27,* and *28*) was included to test the Population IVs students more particularly. Overall the items in this section proved of average difficulty and did tend to discriminate satisfactorily both within and between countries.

Section 4 consisted of simple and short conversations between a man and a woman. Some items like the following test the ability to *interpret* the drift of a conversation:

(Item 35)

Woman: I wonder why everybody should have to take mathematics.
Man: Because everybody needs it, I suppose.

The responses listed were:

*A Bill thinks that mathematics is a useful subject.**
B Ann thinks that mathematics is a useful subject.
C Ann and Bill both think that mathematics is a useful subject.

Section 4 proved more difficult than Section 3. It also discriminated less satisfactorily.

Most of the Listening items in both Populations were point-specific. Items in Section 1 (both Populations), which was concerned with discrimination of sounds, and Section 2 (Population IV), concerned with intonation, were regarded as specific and analytical items. Section 1 also belonged to the category of vocabulary items, while the remaining sections of both Populations were regarded as testing grammatical structures.

Reading Tests. Apart from the items which were based on the comprehension of short sentences or longer texts, the Reading items were regarded as being of the point-specific rather than the global type. For instance, in Population II:

(Item 1)

The students were required to complete the sentence by the use of a word opposite in meaning to the one underscored in the stimulus:

Is your friend a fat boy? No, he is _______.

Response alternatives were:

A *round.*
B *short.*
C *thin.**
D *large.*
E *tall.*

Even more specific were the items in Section 3 which required the student to be able to identify from their reading words which possess a specified sound in common.

(Item 12)

Stimulus word: *even*

Responses:

A *end.*
B *help.*
C *get.*
D *eye.*
E *me.**

Other point specific items refer to grammatical structures as in:

(Item 20)

We have studied English _______ 1964.

A *since**
B *for*
C *from*

The Population IV test, like that for Population II, included items designed to assess the student's ability to recognise the correct pronunciation of the printed words. In Population II it was "sound-letter correspondence" that was tested, while in Population IV "word stress" was the problem. Students were asked to indicate which words had the stress on the same syllables. There were 5 possible answers--none, all, first and second words, first and third words, and second and third words.

(Item 6)

(i) *rea-son-a-ble.*
(ii) *en-ter-tain-ment.*
(iii) *de-liv-er-y.*

A *i and ii are the same.*
B *i and iii are the same.*
C *ii and iii are the same.*
D *all are the same.*
E *none are the same.**

Section 2 like Section 1 of Population IV is a point-specific test where the student was required to identify which of the 5 listed nouns in each item could not be associated with a specified adjective:

(Item 12)

Smart

A *child.*
B *flower.**
C *trick.*
D *people.*
E *dress.*

Although the pretesting gave no indication of more than average difficulty, these items proved very difficult, partly because of the fact that students were required to come to a negative conclusion and partly because the items necessitated a good measure of sensitivity to idiomatic usage. The items had low within-country discrimination, but they tended to discriminate satisfactorily between countries.

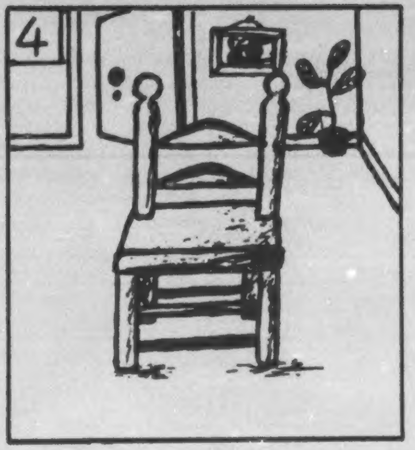
The comprehension items, like those in Population II, were by their nature of the global type, whether they were associated with short sentences or longer texts. The passages which were included to assess the ability of Population IVs students proved very difficult but discriminated satisfactorily in the way it was hoped they would.

The proportions of vocabulary and grammatical structure items were evenly balanced in their contributions to the maximum possible score. Just as the grammatical structure items attempted to cover as many of the basic features of English grammar, like the use of prepositions, tense, number, and different types of pronouns, different aspects of knowledge of vocabulary were also included, like substitution of antonyms in Population II, definitions in Populations II and IV, and collocations of adjectives and nouns in Population IV. The range of difficulty of items in Population IV was greater than in Population II for such reasons as the test inclusion of Population IVs. Some of the passages for reading comprehension in Population IV (*Items 50-60* particularly) were found to be too difficult for large numbers in Population IV, and the percentage of students who failed to reach this group of items is relatively high in many countries. An attempt was made to arrange the categories of items, as well as items within those categories, in order of difficulty. This proved to be fairly successful in both Populations with the exception of *Items 9-16* (Section 2) in Population IV. This is the section where collocations of adjectives and nouns are involved. These items were omitted by as many as 20% to 30% in one country, and in 3 other countries by percentages ranged between 9 and 15. However, these items appeared to discriminate satisfactorily between Population IVs students and the remainder of Population IV.

Speaking Tests. The Speaking test was administered to subsamples of each of the target populations, each student being tested individually. The same test was used for both Populations except that for Population IV the Fluency Section was extended by the addition of free response material. The whole Speaking test consisted of 3 sections--Structural Control, Oral Reading, and Fluency (A and B).

Section 1: Structural Control. This section was made up of 10 tape-recorded questions, statements, or instructions to which the student was required to respond in English, relating his response to the appropriate picture in his test booklet. The following was one of the items:

(Item 4)



The student heard:

"Is there a chair in the room?"

Any 1 of the following 3 responses would be acceptable:

"Yes, there is a chair in the room."
"Yes, there's a chair in the room."
"Yes, there is."

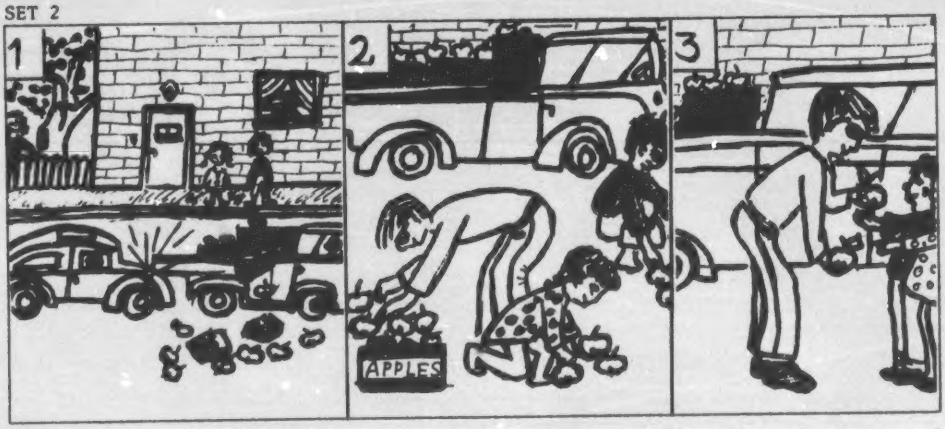
Although in normal conversation yes and no answers would be perfectly proper, for the purposes of the test the student was asked to reply by using complete sentences. In order to avoid lengthy pauses, a student was also asked to reply, "I don't know" if that were in fact the case. All the instructions to the students were in the native language, and his English responses were taped for later scoring. No response, or an unintelligible one, scored zero. A suitable response in which there was an error which did not affect the basis of the test item (for instance, the omission or incorrect use of the article) scored 1, while an appropriate response using the correct verb form with no other types of error scored 2. Scorers were asked not to take account of pronunciation or intonation deviances or hesitancy.

Section 2: Oral Reading. In this section a student was required to read aloud a short passage of simple prose which included conversation. He was told that he would be assessed for the accuracy of his pronunciation and the quality of his phrasing. At the end of the 3 minutes he was allowed for familiarizing himself with the text, he recorded his reading, and this was scored later. The scorers had copies of the printed text in which the features to be scored were indicated. These features included specified vowels and consonants, as well as phrasing and intonational features. In the following extract, the pronunciation items are underscored with a line and the phrasing features by a dotted line. A wrong pronunciation of the underscored word or poor phrasing of the intonation feature lost a point:

He turned to his wife. "What do you think?" he asked.

Section 3: Fluency. In the first part of this section the student was required to describe in English the events set out in a series of pictures. Two sets of pictures were printed, of which the student selected one.

SET 2



The student was informed that his score on this test would be based on the amount and quality of his response (quality being measured by accuracy of grammar and pronunciation), range and suitability of vocabulary, and the variety of grammatical features employed. He was encouraged to say at least 3 sentences about the picture sequence he chose to describe. No time limit was involved, and his response was taped for later scoring.

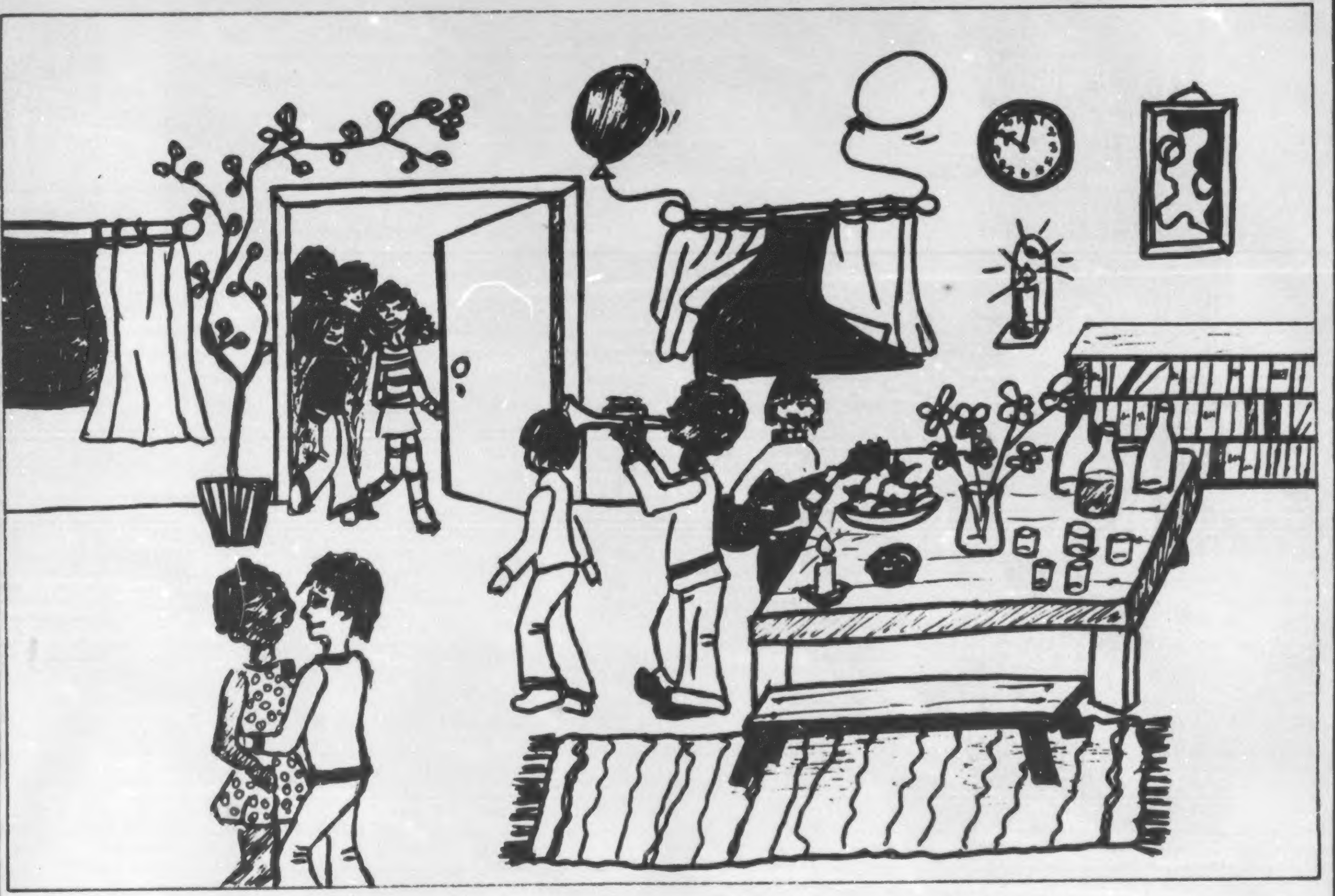
The second part of the Fluency section (set for Population IV only) set out to test the same aspects of English achievement but was intended to be more difficult. One picture was printed in which a party scene was depicted. The student was asked to say something about what had led up to the party, what was occurring at the party, and what might be expected to happen when it was over. He was given the same instructions about the basis of scoring as for the first part of the Fluency test, and 3 minutes were allowed for studying the picture before a response was made. The taped response was scored later, according to the following formula.

Volume. A preliminary examination of the taped responses suggested that 30 intelligible utterances or clauses could be a reasonable maximum. Each clause scored 1.

Accuracy. Three elements were involved, and each element was judged on a 4-point scale:

Grammar
0=3 or more errors or omissions
1=2 errors or omissions
2=1 error or omission
3=No errors or omissions
Vocabulary
0=3 or more incorrect or unintelligible words
1=2 incorrect or unintelligible words
2=1 incorrect or unintelligible word
3=No incorrect or unintelligible words
Pronunciation
0=3 or more errors
1=2 errors
2=1 error
3=No errors

Variety. After the Accuracy of the response had been scored, the tape was played a third time and, according to a 4-point scale, the scorers made a global assessment of the Variety of Tense and Grammatical Structures employed and the Range of Vocabulary. So far as the latter was concerned, a basic list of 6 words was established for the 2 sequences in Fluency A and of 10 words for the 1 picture in Fluency B.



The scoring of the Variety of Tense and Grammatical Structures was more subjective than that of Vocabulary:

0=No variety of either tense or structures
1=Some variety of either tense or structures
2=Some variety in both tense and structures
3=Considerable variety in both tense and structures

The Range of Vocabulary was scored as follows:

0=Vocabulary less than the basic word list
1=Vocabulary limited to the basic word list
2=Vocabulary exceeding the basic word list but to a limited degree
3=Vocabulary exceeding the basic word list to a major degree

Achieving the Total Score. A final weighting scheme was prepared at Educational Testing Service by John B. Carroll for the whole of the Speaking test (Populations II and IV). This was determined on the basis of 1,186 cases after eliminating those with missing data on any part of the Speaking tests or with zero scores on the Volume variable (number of intelligible clauses). In the first stage of the analysis, the problem was to develop weighting formulas from compositing the elements of the Fluency tests. The Relative Quality Scores were formed by dividing the "Quantity" of correct scores by the "Volume" (variable 4 below). Thus, the initial variables involved were as follows:

1. Global Rating [scaled 0 (low) to 4 (high), given over all aspects of the Speaking test, including Structural Control and Oral Reading]
4. Number of Intelligible Clauses, Fluency A (B)
5. Accuracy 1, Grammar, Fluency A (B)
6. Accuracy 2, Vocabulary, Fluency A (B)
7. Accuracy 3, Pronunciation, Fluency A (B)
8. Relative Accuracy, Grammar, Fluency A (B)
9. Relative Accuracy, Vocabulary, Fluency A (B)
10. Relative Accuracy, Pronunciation, Fluency A (B)
11. Variety 1, Tense and Structure, Fluency A (B)
12. Variety 2, Range of Vocabulary, Fluency A (B)

The first stage analysis yielded 3 composites, as follows:

13. Quantity Score, variables 5, 6, 7 (or 17, 18, 19), 1, 1, 1, respectively
14. Quality Score, variables 8, 9, 10 (or 20, 21, 22), 3, 1, 2, respectively
15. Variety Score, variables 11, 12 (or 23, 24), 2, 1, respectively

It was noted, incidentally, that variables 5, 6, and 7 were highly co-linear, with the result that their beta-weights varied widely through sampling error; for this reason they were weighted equally.

In the second stage of the analysis, the variables considered were variables 13, 14, 15 (or 25, 26, 27), plus variables 2 (Total Score, Structural Control) and 3 (Total Score, Oral Reading). In these analyses, variable 2 usually had significant positive beta-weights, whereas variable 3 had either non-significant weights or significantly negative beta-weights. It was therefore decided to drop variable 3 from the final composites altogether. Variables 13, 14, 15 (or 25, 26, 27) nearly always had highly significant positive beta-weights. With account taken of average relativized beta-weights, the integer weights for composing a final Total Speaking Score were as follows:

Variable 2 (Structural Control):	8
Variable 3 (Oral Reading):	0
Variable 13 (Quantity):	1
Variable 14 (Quality):	6

Variable 15 (Variety): 15

The final weighting formulas were adjusted so that the Quantity, Quality, Variety, and Total Speaking Scores would be approximately comparable for all Populations, i.e., regardless of whether only Fluency A Scores, or both Fluency A and B Scores, were present. (In effect, the A and B Scores are averaged for Population IV).

Tables 5 and 6 present selected correlations among the global ratings, the initial variables, and the final composites, for all samples from Population II and IV. The validity of the Total Speaking Score is .859 for the first set of pooled samples and .734 for the second set.

Table 5

Selected Intercorrelations of Initial Variables, Composites, and Global Ratings, with Means and Standard Deviations of all Variables

N=875 cases from Population II

Variable	$\overline{X}$	SD	1	13	14	15	31
1. Global Ratings	1.66	1.01	1.000	.724	.486	.771	.859
2. Total Structural Control	8.16	5.04	.670	.419	.475	.558	.822
3. Total Oral Reading	25.32	6.46	.402	.243	.295	.340	.448
4. Fluency A, No. Clauses	6.51	4.03	.671	.979	.232	.628	.726
5. Fluency A, Accuracy 1 Grammar	15.88	11.13	.720	.986	.401	.663	.798
6. Fluency A, Accuracy 2 Vocabulary	17.20	11.37	.697	.986	.315	.633	.750
7. Fluency A, Accuracy 3 Pronunciation	15.99	10.35	.719	.976	.415	.692	.810
8. Fluency A, Relative Accuracy: Grammar	2.36	0.52	.442	.372	.876	.388	.567
9. Fluency A, Relative Accuracy: Vocabulary	2.62	0.40	.279	.212	.508	.192	.292
10. Fluency A, Relative Accuracy: Pronunciation	2.43	0.53	.324	.223	.744	.301	.436
11. Fluency A, Variety 1 Tense and Structure	0.88	0.81	.729	.610	.417	.948	.811
12. Fluency A, Variety 2 Range of Vocabulary	0.69	0.81	.603	.578	.301	.777	.686
13. Fluency A, Quantity	49.07	32.29	.724	1.000	.382	.673	.798
14. Fluency A, Quality	87.22	14.20	.486	.382	1.000	.424	.621
15. Fluency A, Variety	36.82	32.47	.771	.673	.424	1.000	.864
31. Total Speaking	238.36	95.76	.859	.798	.621	.864	1.000

The inter-rater reliabilities of the global ratings were .754 for Writing and .787 for Speaking. These correlations are quite respectable considering that they are for 2 raters; they may be taken as estimates of the reliability of a single rating.

The inter-rater reliabilities are quite high whenever they are based on Quantity; they are lower when based on Quality or Variety. Nevertheless, all the final composite scores may be considered to have a sufficient degree of reliability for group comparisons and the other analyses performed in this project.

Writing Tests. The same Writing test was used for both Populations. It was a group test composed of 4 sections, the first 3 calling for "guided," closely structured responses, while the fourth was a "free-writing" composition. The time limit

Table 6

Selected Intercorrelations of Initial Variables, Composites, and Global Ratings, with Means and Standard Deviations of all Variables

N=311 cases from Population IV

Variable	$\overline{X}$	SD	1	28	29	30	31
1. Global Rating	2.13	0.90	1.000	.693	.472	.645	.734
2. Total Structural Control	11.65	5.67	.439	.327	.471	.406	.743
3. Total Oral Reading	28.82	9.31	.204	.236	.153	.157	.437
Fluency A							
4. No. Clauses	9.56	5.88	.621	.884	.246	.574	.688
5. Accuracy: Grammar	25.09	16.79	.642	.881	.348	.592	.719
6. Accuracy: Vocabulary	25.96	16.51	.627	.876	.279	.581	.693
7. Accuracy: Pronunciation	24.44	16.00	.643	.902	.383	.624	.755
8. Relative Accuracy: Grammar	2.56	0.41	.338	.278	.725	.304	.412
9. Relative Accuracy: Vocabulary	2.69	0.27	.184	.130	.370	.196	.219
10. Relative Accuracy: Pronunciation	2.51	0.45	.310	.281	.686	.338	.439
11. Variety, Tense and Structure	1.18	0.77	.513	.520	.358	.800	.624
12. Variety, Range of Vocabulary	1.11	0.85	.344	.516	.374	.644	.601
13. Quantity	75.49	48.80	.644	.895	.340	.605	.729
14. Quality	92.46	10.79	.415	.351	.899	.408	.537
15. Variety	52.04	30.64	.529	.605	.424	.869	.719
Fluency B							
16. No. Clauses	12.48	7.01	.605	.929	.324	.688	.790
17. Accuracy: Grammar	33.15	19.83	.627	.927	.398	.691	.808
18. Accuracy: Vocabulary	34.56	19.78	.625	.931	.361	.697	.801
19. Accuracy: Pronunciation	32.87	19.93	.612	.917	.416	.701	.810
20. Relative Accuracy: Grammar	2.62	0.27	.328	.276	.630	.238	.374
21. Relative Accuracy: Vocabulary	2.76	0.21	.235	.148	.354	.169	.200
22. Relative Accuracy: Pronunciation	2.58	0.35	.291	.281	.684	.356	.424
23. Variety, Tense and Structure	1.41	0.76	.550	.525	.394	.807	.628
24. Variety, Range of Vocabulary	0.58	0.94	.439	.597	.180	.649	.590
25. Quantity	100.59	59.20	.625	.930	.394	.700	.811
26. Quality	94.66	7.90	.391	.340	.801	.360	.484
27. Variety	51.08	31.54	.596	.648	.366	.877	.720
28. Total Quantity	88.04	49.38	.693	1.000	.404	.719	.847
29. Total Quality	93.56	8.01	.472	.404	1.000	.452	.600
30. Total Variety	51.56	27.13	.645	.719	.452	1.000	.824
31. Total Speaking	326.33	102.71	.734	.847	.600	.824	1.000

for the first 3 sections was 15 minutes, but those who completed the sections in less than the allocated time were advised to proceed to the fourth section, the time limit for which was 25 minutes. All test instructions were given in the student's native language.

Section 1 consisted of 13 sentence-completion items. There was no concern for testing students' ability to make correct use of a variety of grammatical features--prepositions, relative pronouns, adverbs, and tense forms. The scoring of this section was simply right or wrong, each correct answer scoring 1 point, and an incorrect response scoring zero. The items were also to a certain extent open ended in the sense that the student was not offered alternative responses but had to produce his own correct completion of a sentence. The 2 examples preceding the actual test were:

(i) Do you know when America was discovered?
(ii) Denis is leaving the city by train.

Section 2 was divided into 4 categories, and here the student was required in some cases to produce 2 or 3 words to complete a given sentence. The 4 sections involved the correct use of pronouns:

(Item 14)

Isn't your brother going out?
No, he isn't going out.

or the ability to make correct changes in the number of nouns and verbs:

(Items 16 and 17)

She saw a policeman in that car.
She saw some policemen in those cars.

or changes in the tense of verbs:

(Item 20)

I make a cake every Sunday.
I made a cake last Sunday.

or changes in voice:

(Item 22)

The policeman asked Bill his name.
Bill was asked his name by the policeman.

Here again the scoring was on a right or wrong basis.

Section 3 contained 8 sentence completion items of a multiple choice type. The section was concerned with correct word order, and the students were given a sentence stem and 3 words or phrases arranged in random order with which to complete the sentence. As for the first 3 sections, completely objective scoring was possible, the students being required to mark the alternative word order they selected.

(Item 24)

These shoes are ______ ______ ______
(i) enough (ii) for me (iii) big.

A i, iii, ii.
B ii, i, iii.

C ii, iii, i.
D iii, i, ii.
E iii, ii, i.

Section 4. In this section the students were required to produce a written composition of not more than 200 words on the subject of travel. The only limit imposed on the free-response of the students was the requirement that they should use 12 given words in any order and in any grammatical relationship. They were told that the scoring would take account of spelling and grammatical accuracy, the number of intelligible statements that were written, the range of vocabulary, and the variety of tense forms and grammatical structures. In addition the students were asked to try to produce a "cohesive" account rather than a series of unconnected statements.

The scoring system used for the composition section of the Writing test was developed from that originally devised to score spoken fluency. Each composition was scored for Volume, Accuracy, Variety, and Cohesiveness, as follows.

Volume. The scorer counted the number of intelligible clauses written, up to a maximum of 200 words, marking the end of each intelligible clause with an oblique line. Any unintelligible material was ignored. If a student had failed to finish copying out the final version of his composition, but had nevertheless completed the composition in draft form, the draft was retained and was used in scoring from the point where the final version had been left unfinished. If the final version of the composition was complete, the rough draft was destroyed.

Accuracy. Each intelligible clause was then scored for Accuracy on a 4-point scale within each of the following 3 categories: grammar, vocabulary, spelling. The rating-scales used in each category are shown below:

Grammar
0=3 or more errors or omissions
1=2 errors or omissions
2=1 error or omission
3=No errors or omissions
Vocabulary
0=3 or more incorrect or unintelligible words
1=2 incorrect or unintelligible words
2=1 incorrect or unintelligible word
3=No incorrect or unintelligible words
Spelling
0=3 or more spelling errors
1=2 spelling errors
2=1 spelling error
3=No spelling errors

Variety. After the student's composition had been scored for Volume and for Accuracy, as detailed above, the scorer gave a global rating for the Variety of the writing as a whole, using a 4-point scale within each of the following 2 categories: Variety of Tense and Structure and Range of Vocabulary. The rating-scales used in each of these categories are shown below:

Variety of Tense and Structure
0=No variety in either tense or structure
1=Some variety in either tense or structure
2=Some variety in both tense and structure
3=Considerable variety in both tense and structure

Range of Vocabulary. The criterion for Range of Vocabulary was established by reference to the basic list of 12 words which the students had been instructed to include in their written composition. Rating was as follows:

0=Vocabulary less than the basic word-list
1=Vocabulary limited to the basic word-list

2=Vocabulary exceeding the basic word-list, but to a minor extent
3=Vocabulary exceeding the basic word-list to a major extent

Cohesiveness. After rating the composition for its Variety, the scorer then rated the student's writing as a whole for its Cohesiveness, using the following 4-point scale:

0=No attempt to link statements
1=Minimal linking of statements (e.g., the repeated use of "and")
2=Some variety in the linking of statements (e.g., "and," "but," "as well as," etc.)
3=Use of subordinating conjunctions and embedded statements

When the student's script had been scored as described above, the scorer made a global assessment of the Quality of the student's composition as a whole and assigned him a rating on a 5-point scale, defined as follows:

0=Extremely poor response
1=Poor response
2=Reasonable response
3=Good response
4=Extremely good response

The assignment of a global rating completed the scoring of the Writing test.

Achieving the Total Score. As was the case with the Speaking test, Professor Carroll devised a weighting formula for the whole of the Writing test, and the following reflects his conclusions.

English Writing Test. The scoring system for the total English Writing test was determined on the basis of 8,049 cases. Regression analyses were performed separately for each of the samples taken from several countries and for all samples pooled.

The initial variables used in this analysis were as follows:

1. Global Rating [scale 0 (low) to 4 (high)]
2. Volume: Number of Intelligible Clauses
3. Accuracy 1: Grammar
4. Accuracy 2: Vocabulary
5. Accuracy 3: Spelling
9. Variety of Tense and Structure
10. Range of Vocabulary
11. Cohesiveness

In the first stage of the analysis, the interest was in determining whether several somewhat independent composites could be formed from these variables. It was noted that the Accuracy Scores (variables 3, 4, and 5) would be regarded as measures of "Quantity of Correct Writing," since they represented sums of points (on a 0-3/?/ scale) to each intelligible clause. If these scores were each divided by variable 2 (the number of intelligible clauses), the resulting scores could be regarded as measures of relative Quality of writing, since they would represent average points awarded per clause. Therefore, 3 new variables were established:

6. Average Accuracy: Grammar=Variable 3)/(Variable 2)
7. Average Accuracy: Vocabulary=(Variable 4)/(Variable 2)
8. Average Accuracy: Spelling=(Variable 5)/(Variable 2)

By means of a regression analysis procedure where relativised regression weights were averaged, it was decided to form 3 composites:

Quality Score: Weights of 2, 1, 1 for variables 3, 4, 5, respectively
Quantity Score: Weights of 2, 1, 1 for variables 6, 7, 8, respectively
Variety Score: Weights of 2, 1, 2 for variables 9, 10, 11, respectively

In the pooled samples, the correlations of these variables among themselves and with the global ratings were as follows (*N*=8,049):

	Quantity	*Quality*	*Variety*	*Global Rating*
12. Quantity Score	1.000	.546	.717	.729
13. Quality Score	.546	1.000	.502	.537
14. Variety Score	.717	.502	1.000	.788

As will be seen, these 3 scores make independent and significant contributions to the prediction of the global rating, and it may be concluded that they measure different aspects of writing performance.

At the second stage of analysis, the problem was to form a Total Writing Score composite from these 3 composite variables. Separate regression analyses were made for each Population-country combination. In these analyses, the beta-weights for all 3 composites were uniformly positive; those for variable 12 (Quantity) and variable 14 (Variety) were in every case significant at the 1% level, and those for variable 13 (Quality) were significant at that level in 3 of the 7 cases. (This is partly a matter of different sample sizes.) Appropriate integer weights for the 3 variables, allowing for their different beta-weights and variances, were determined to be 1, 10, and 20, respectively. In the pooled sample, the beta-weights for these 3 composites were .293, .115, and .520, respectively, with a multiple correlation of .828.

Table 7 shows selected correlations of global ratings, initial variables, and composite scores. The Total Writing Score, as weighted by the formula established, correlates .819 with the global ratings in the pooled sample.

Table 7

Selected Correlations of Global Ratings, Initial Variables, and Composites Formed for English Writing

N=8,049

Variable	$\bar{X}$	SD	*Correlations* 1	12	13	14	15
1. Global Ratings	1.76	1.00	1.000	.729	.537	.788	.819
2. No. Intelligible Clauses	17.28	7.85	.687	.979	.409	.684	.902
3. Accuracy 1: Grammar	43.00	21.90	.723	.993	.575	.716	.937
4. Accuracy 2: Vocabulary	46.30	22.05	.719	.988	.500	.704	.923
5. Accuracy 3: Spelling	46.59	22.80	.721	.985	.514	.702	.921
6. Relative Accuracy: Grammar	2.42	0.40	.456	.497	.895	.457	.559
7. Relative Accuracy: Vocabulary	2.65	0.28	.332	.296	.604	.279	.341
8. Relative Accuracy: Spelling	2.65	0.37	.386	.354	.630	.320	.395
9. Tense and Structure	1.45	0.82	.688	.587	.436	.830	.757
10. Range of Vocabulary	0.71	1.05	.467	.428	.276	.625	.554
11. Cohesiveness	1.67	0.93	.673	.644	.443	.873	.804
12. Quantity Score	178.89	87.72	.729	1.000	.546	.717	.939
13. Quality Score	101.46	11.25	.537	.546	1.000	.502	.616
14. Variety Score	139.10	72.71	.788	.717	.502	1.000	.909
15. Total Writing	419.45	155.42	.819	.939	.616	.909	1.000

Although the univariate statistics give evidence of considerable skewness in some of the variables, inspection of certain relevant scatterplots of means of variables with mean global ratings over samples gave no indication that it would be worthwhile to carry out non-linear transformations on some of these variables. It was in any case thought best to leave the variables in their original forms, so that any desired transformations could be applied later.

A word should be said about the assumptions underlying the procedure of determining weights from regression analyses against global ratings as criteria. The assumption is that global ratings themselves are too much influenced by transitory subjective factors (such as variations in the perceived Average Quality of the material being judged) to be used as criterion variables, but that over the total sample of cases, the global ratings do provide a measure--though an unreliable one--of Quality. Thus, an optimal weighting of detailed scores based on their regression weights in predicting global weightings would provide a sufficiently reliable and quasi-objective set of criterion scores. It is assumed also that the detailed scores are less subject to the transitory subjective factors that would affect the global ratings.[4]

Difficulties of Preparing and Administering Speaking Tests. Many of the students taking part in the IEA study had never been exposed to oral testing, and the English Committee decided, in order to minimize the disadvantages from which they might suffer, that students should be given ample opportunity to work through untimed practice items with their teachers, who would be encouraged to prompt the students as might be necessary during the practice period. All test instructions were given in the native language.

It is difficult to provide unambiguous pictorial material, and the experience of the English Committee verified this. The original material had to be discarded because of frequent misinterpretation. Finally, the black and white sketches which were used were made as simple and as austere as was compatible with the needs of the item. Pictorial material connected with free composition was less difficult to devise than that which was needed for testing grammatical structures.

Some difficulty was experienced in the timing of pauses between elements in the stimulus questions. If the pauses are too short, able students attempting a complex and possibly long response will not be able to complete their answer. If the pause is too long, students become confused and attempt to elaborate a correct, though possibly short, response.

Possibly the greatest difficulty arises from the scoring of open ended and free composition responses. Because of the possibly subjective scoring problems which arise in such cases, it was decided to employ a small team of experienced native speakers of English for all free-response items. As a check on the stability of scoring standards, each member of the team was given periodically a random sample of tapes and scoring scripts scored by other members of the team, as well as a random sample of material he himself had scored previously.

RELIABILITIES OF TESTS

The validity of the measures was ensured, so far as possible, by the consultations which were conducted with National Centres at all stages. There is reason, therefore, to be confident that effective measures were produced; but until it is possible to compare the performance of students on the present tests with their performance on others, such as external school examinations in English, statistical evidence of the validity of the cognitive tests cannot be provided. Another qualification must be recorded: the usefulness of the measures may have been affected by the need to translate some of the instructions, as well as the responses, in several instances. There is also the question of the reliabilities of the Subscores, since there were very few items in each category or group. Reliability data are reported for Total Scores and Subscores in all aspects of English tested and for all countries. However, in some instances, though a country might have participated in testing a particular aspect of English, the sample of students for some parts of the test (Listening, for instance) was so small, or the number who completed the test was so unrepresentative of the total sample, that a computation of reliability became pointless.

Estimates of the reliabilities using Kuder-Richardson Formula 20 (Kuder and Richardson, 1937; see also Lord and Novick, 1968) were obtained for each Population tested in each country for the Reading and Listening tests. These are presented in Tables 8 and 9. This estimate of reliability is based on the internal consistency of all the items in the total test or subsection of the test, as indicated by the heading of the column in which the various reliabilities appear. The Subscores of the test involve fewer items, and the reliabilities might therefore be expected to be lower, which is generally the case. The estimated reliabilities for the non-cognitive measures were calculated by using the Coefficient α (Cronbach, 1951). Separate coefficients were calculated for each test and for each country. It was not possible to produce computations of the reliabilities of any of the tests in Speaking and Writing.

Generally speaking the achievement tests in Reading, and to a lesser extent in Listening (in those countries where the conditions of testing justified the computation of reliabilities), proved to be relatively sound measures when taken in each case as 1 complete test. Since they are based on fewer items, the sectional scores in Reading and Listening tend to be far less reliable.

Listening

Table 8

Means, Standard Deviations, and Reliabilities for Listening Total Test and Subscores--Population II*

	Total Test			Subscores in Order of Sections					
				I			II		
Country	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.
Belgium	9.4	5.7	.76	5.5	3.1	.50	3.9	3.7	.64
Finland	8.1	6.2	.96	5.0	3.1	.64	3.1	4.1	.71
Italy	8.9	6.4	.83	5.1	3.2	.59	3.8	4.1	.77
Sweden	15.5	4.6	.95	8.0	2.2	.35	7.6	3.2	.64
Median Value	9.1	5.9		5.3	3.1		3.8	3.9	
No. of Items		24			12			12	

*Total Test contains information based only on Subsections 1 and 2, the 2 multiple choice tests, and excluded Subsection 3, the Dictation test.

Subscores:
Section I: Discrimination of Sounds
Section II: Listening Comprehension

Table 9

Means, Standard Deviations, and Reliabilities for Listening Total Test and Subscores--Population IV

				Subscores in Order of Sections											
Country	Total Test			I			II			III			IV		
	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.
Belgium	16.9	7.0	.76	7.1	2.7	.49	5.8	1.9	.55	2.4	2.7	.55	1.7	2.6	.53
Chile	7.1	4.5	.67	2.9	2.8	.42	4.3	2.1	.56	0.2	1.9	.57	-0.2	1.8	.12
Finland	23.2	6.1	.71	8.4	2.3	.38	5.5	2.0	.51	4.2	2.4	.50	5.0	2.3	.52
Hungary	16.3	7.3	.75	6.2	3.0	.52	6.2	1.6	.45	2.1	2.8	.54	1.8	2.8	.54
Italy	14.7	7.6	.84	5.7	2.8	.50	5.8	1.6	.48	2.1	2.6	.55	1.6	2.7	.62
Sweden	26.2	5.4	.78	9.2	1.9	.32	5.4	1.9	.50	5.4	2.1	.46	6.5	1.8	.52
Median Value	16.6	6.5		6.6	2.7		5.6	1.9		2.2	2.5		1.7	2.4	
No. of Items		36			12			8			8			8	

Subscores:

Section I:	Discrimination of Sounds
Section II:	Intonation
Section III:	Listening Comprehension
Section IV:	Listening Comprehension (Conversation)

Reading

The reliabilities of the total scores are relatively high in both Populations but higher in Population II than in Population IV. There are only small differences between countries. The Subscores have almost invariably lower reliability coefficients than the Total Scores, and this is true of all countries. The sections with the highest reliability levels are Section VI (Comprehension of Longer Passages) in Population II, and Section V in Population IV. The next highest level of reliability in Population II is obtained by Section V (Comprehension of Short Sentences). It is to be expected that the countries with the lowest or highest reliabilities on the Subscores tend to have the lowest or highest reliabilities on the Total Scores also. The least reliable of the sectional Subscores in Population II are in respect

Table 10

Means, Standard Deviations, and Reliabilities for Reading Total Test and Subscores–Population II

	Total Test			Subscores in Order of Section																	
				I			II			III			IV			V			VI		
Country	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.
Belgium (Fr.)	22.8	15.2	.91	1.6	1.8	.50	4.1	2.4	.63	2.2	2.8	.45	2.7	2.3	.58	5.6	4.8	.78	6.6	4.7	.80
FRG	29.9	16.2	.93	3.0	1.8	.56	5.0	2.3	.67	3.9	3.0	.51	2.8	2.5	.62	7.1	4.9	.78	8.3	4.9	.82
Finland	16.7	17.8	.93	1.6	2.1	.64	2.7	2.7	.72	2.3	3.2	.57	1.4	2.6	.62	4.0	5.1	.77	4.8	5.0	.78
Israel	26.3	18.7	.95	3.0	2.2	.73	2.9	2.5	.69	3.7	3.3	.62	3.2	2.7	.74	6.8	5.3	.83	7.3	5.9	.88
Italy	18.6	17.3	.94	1.7	2.2	.71	3.1	2.6	.71	3.1	3.5	.68	2.0	2.6	.70	4.1	4.7	.80	4.9	5.3	.86
Netherlands	37.4	16.4	.94	3.5	2.2	.69	5.2	2.4	.73	5.5	3.2	.66	4.4	2.2	.71	8.2	4.2	.77	10.5	4.6	.83
Sweden	35.0	14.9	.93	2.7	2.0	.58	4.9	2.4	.66	5.3	2.7	.52	4.5	2.0	.66	8.1	4.3	.76	9.6	4.5	.80
Thailand	18.4	13.4	.87	2.1	1.8	.53	2.3	2.1	.52	2.9	3.1	.59	2.2	2.3	.59	4.6	4.0	.65	4.3	3.9	.70
Median Value	24.5	16.3		2.4	2.1		3.6	2.4		3.4	3.1		2.7	2.4		6.2	4.7		6.9	4.8	
Number of Items		60			6			8			10			6			14			16	

Subscores:

Section I:	Recognition of Antonyms
Section II:	Letter-Sound Correspondence
Section III:	Recognition of Structural Features
Section IV:	Vocabulary Recognition
Section V:	Comprehension of Short Sentences
Section VI:	Comprehension of Longer Passages

Table 11

Means, Standard Deviations, and Reliabilities for Reading Total Test and Subscores–Population IV

	Total Test			Subscores in Order of Section														
				I			II			III			IV			V		
Country	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.	Mean	S.D.	R.
Belgium (Fr.)	24.2	10.8	.86	1.5	2.0	.40	1.1	1.8	.46	3.0	2.5	.41	6.4	2.8	.58	12.3	5.7	.82
Chile	10.6	9.8	.87	1.6	2.0	.41	0.7	1.5	.34	0.9	2.2	.29	2.9	3.0	.58	4.9	5.2	.81
FRG	41.2	6.1	.68	5.1	1.8	.49	2.8	1.7	.31	5.1	1.8	.27	8.7	1.5	.17	19.5	3.0	.57
Finland	35.6	8.3	.78	2.6	2.3	.51	2.9	2.0	.39	5.5	1.9	.33	8.7	1.7	.28	15.9	4.5	.72
Hungary	20.1	11.2	.85	1.8	2.1	.44	1.6	2.2	.52	2.9	2.8	.57	5.3	2.7	.51	8.6	5.5	.77
Israel	27.4	12.9	.91	3.0	2.4	.58	1.8	2.2	.59	4.6	2.5	.56	7.2	2.8	.66	11.4	6.7	.87
Italy	20.2	12.6	.90	2.2	2.4	.55	1.7	1.9	.59	2.8	2.6	.41	4.9	3.0	.61	8.8	7.0	.90
Netherlands	42.9	6.5	.69	5.0	2.2	.59	3.9	1.9	.24	5.4	2.1	.39	9.0	1.5	.14	19.7	3.0	.59
Sweden	39.5	9.0	.84	3.8	2.6	.66	3.3	1.7	.26	5.2	2.2	.46	9.2	1.8	.41	17.8	4.5	.79
Thailand	19.7	9.1	.76	0.9	1.7	.67	1.2	1.7	.17	3.5	2.6	.47	5.7	2.7	.46	8.4	4.7	.66
Median Value	25.8	9.4		2.4	2.1		1.7	1.8		4.0	2.3		6.8	2.7		11.8	4.9	
Number of Items		60			8			8			8			11			25	

Subscores:

Section I:	Recognition of Vocabulary
Section II:	Collocations
Section III:	Recognition of Grammatical Structures
Section IV:	Comprehension of Short Sentences
Section V:	Comprehension of Longer Passages

of Section III (Recognition of Structures) and in Population IV the least reliable subsection appears to be Section II (Collocations). There is a marked uniformity between countries in respect of which Subscores are most and least reliable.

The reliabilities of Total Scores tend to be high in Population IV also, but lower than in Population II. In Population IV the only subsection emerging with fairly high levels of reliability is Section V (Comprehension of Longer Passages), as in Population II.

Comparison of Categories of Items--Specific/Global and Vocabulary/Structural

The items which are included in the 2 categories A and B in Table 12 cut across the subsections discussed already. Considerable caution should be exercised in comparing Columns 1 with 2 of A and 1 with 2 of B in either Population, as well as in comparing the results of the 2 Populations, since in the first place the items are few and secondly, although the types of items in Population II and Population IV are identical, the items themselves are different. The several categories tend to have lower reliabilities in Population IV than in Population II. The difference in level of reliabilities between "analytical (specific)" tests and "global" tests is greater in Population IV than in Population II, and this may be due to there being fewer analytical tests than global tests in Population IV. In general, in both Populations the global tests tend to be more reliable than analytical tests. A comparison of Vocabulary and Structure items points to similar conclusions--in both cases the level of reliability is lower in Population IV than in Population II, though the differences between the 2 Populations are greater in respect of vocabulary type tests than structural items. There also appears to be less consistency between countries in reliability levels for Population IV than for Population II. This may arise from the students in Population IV being a more heterogeneous sample, in respect of length of English studies and type of English programme, than the sample of students in Population II. The most reliable type of test in the present investigation appears to be that which highlights problems of grammatical structure. Within the alternative dichotomy, analytical and global tests, the latter is invariably the more reliable (except in Thailand in Population II).

Table 12

Reliabilities (Kuder-Richardson 20) of the English Reading Subtests Pertaining to the Two Divisions (A and B) of the Total Test for Each Population in Each Country

	Population II			
	A		B	
Country	Analytical	Global	Vocabulary	Structure
Belgium	.82	.87	.78	.90
Chile	-	-	-	-
FRG	.82	.88	.79	.90
Finland	.87	.88	.85	.91
Hungary	-	-	-	-
Israel	.89	.92	.86	.94
Italy	.89	.90	.87	.93

Table 12 *(Continued)*

Netherlands	.89	.89	.87	.92
Sweden	.82	.85	.79	.89
Thailand	.83	.81	.77	.87
Median Value	.85	.88	.82	.91
No. of Items	30	30	20	40
	Population IV			
	A		B	
	Analytical	Global	Vocabulary	Structure
Belgium	.66	.84	.61	.85
Chile	.69	.89	.68	.88
FRG	.54	.57	.48	.61
Finland	.58	.70	.54	.72
Hungary	.67	.80	.56	.84
Israel	.76	.89	.66	.90
Italy	.61	.91	.55	.90
Netherlands	.57	.58	.49	.62
Sweden	.65	.72	.61	.74
Thailand	.44	.73	.23	.77
Median Value	.63	.77	.56	.81
No. of Items	24	36	16	44

SCORE DISTRIBUTIONS

The distribution of scores in the Reading, Listening, Speaking, and Writing tests are set out in Tables 13-15. These reflect both the inherent characteristics of the tests as well as the abilities in English of the samples of students who were tested. The nature of the IEA study, drawing as it did on samples of students from countries differing considerably in their organisation of English in the curriculum at every stage, and on samples of students who varied in their home background and their exposure to English outside school, as well as in experience of English in school, made it inevitable that the test results would have wide distribution frequencies ranging from poor to very good. The number of intervals of distribution vary according to the aspect of English that is tested--Reading, Listening, Speaking, or Writing--but in all cases they are in terms of the original, uncorrected scores.

The Population II and IV Listening tests served their purpose satisfactorily in distributing the scores widely within each country. At the same time, in most countries where a sufficient sample of students attempted or completed the test, the latter appear to have a level of difficulty which is appropriate to the Population being tested, though this is more particularly the case with regard to Population II.

Listening

Table 13

Table of % Frequencies for Total Listening Populations II and IV

Country/Population	3	6	9	12	15	18	21	24	27	30	33	Sum Wts.
Population II												
Belgium	0.3	2.4	7.5	15.1	13.8	13.9	16.4	10.8	10.1	6.2	3.4	671
FRG	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	6
Finland	1.6	9.5	20.3	15.5	9.5	8.1	8.6	9.2	8.5	6.3	2.9	2118
Israel	75.1	18.0	6.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	17
Italy	2.0	11.8	21.4	24.3	19.6	10.8	6.5	2.3	0.4	0.7	0.2	694
Netherlands	62.5	37.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	5
Sweden	0.1	0.7	2.5	3.5	7.2	8.4	10.7	14.3	21.4	21.5	9.8	1450
Thailand	64.4	17.8	17.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	12
Population IV												
Belgium	0.0	0.5	1.0	3.2	11.6	19.2	23.2	18.3	13.2	6.9	2.8	1394
Chile	5.5	0.9	6.6	26.5	38.4	14.7	6.3	1.2	0.0	0.0	0.0	231
FRG	99.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1
Finland	0.0	0.0	0.0	0.3	2.0	5.2	12.8	21.4	26.7	23.0	8.5	2296
Hungary	0.0	0.1	0.5	2.9	13.3	21.9	21.6	15.9	13.8	8.1	2.0	1063
Israel	47.3	24.1	12.3	8.1	8.1	0.0	0.0	0.0	0.0	0.0	0.0	10
Italy	3.3	1.7	1.1	5.2	16.7	20.7	18.1	13.0	12.8	7.3	0.1	314
Netherlands	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2
Sweden	0.2	0.0	0.4	0.5	0.5	1.8	7.2	13.2	28.3	34.6	13.3	997
Thailand	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1

Reading

Table 14

Table of % Frequencies for Total Reading Populations II and IV

Country/ Population	0	5	10	15	20	25	30	35	40	45	50	55	Sum Wts.
Population II													
Belgium	0.0	0.7	4.6	11.8	12.7	13.7	14.4	12.0	13.1	11.6	5.4	1.1	687
FRG	0.2	0.8	1.4	4.1	10.8	12.8	9.7	12.8	11.4	13.9	15.8	6.2	1074
Finland	0.1	0.6	10.2	22.6	15.4	8.2	9.0	9.5	8.5	8.1	5.7	2.0	2100
Israel	1.1	1.3	5.8	10.9	10.1	8.7	9.0	10.2	11.0	12.8	10.7	8.1	1065
Italy	1.6	3.1	7.6	17.1	17.3	12.7	10.0	6.5	6.5	6.0	7.7	3.7	791
Netherlands	0.0	0.1	1.0	3.6	6.2	6.1	6.2	8.2	9.4	15.3	26.0	18.1	2090
Sweden	0.3	0.4	1.1	2.5	4.6	7.2	8.4	13.6	15.8	17.3	20.0	8.9	1477
Thailand	0.0	0.3	3.4	11.1	18.2	18.7	18.1	11.9	10.3	5.2	2.1	0.6	1951
Population IV													
Belgium	0.2	0.3	1.5	6.2	10.3	17.6	21.4	23.2	11.6	5.4	2.0	0.3	1440
Chile	2.1	3.6	14.6	32.4	23.3	12.4	4.9	2.9	1.5	1.3	0.9	0.2	2118
FRG	0.0	0.1	0.0	0.0	0.0	0.5	1.5	7.1	26.0	43.4	20.6	0.8	1374
Finland	0.0	0.0	0.0	0.0	0.8	3.3	10.7	22.1	31.6	22.1	8.6	0.9	2310
Hungary	0.0	0.2	0.9	8.9	21.5	22.0	16.7	15.3	9.9	4.0	0.5	0.2	1063
Israel	1.2	0.9	2.2	5.2	7.6	12.0	18.0	20.1	16.5	11.0	4.1	1.3	604
Italy	0.4	0.0	4.1	15.0	21.7	15.6	12.3	8.9	14.3	5.8	1.9	0.0	324
Netherlands	0.0	0.0	0.0	0.1	0.1	0.2	0.7	5.0	22.0	39.0	27.8	5.1	1561
Sweden	0.1	0.0	0.2	0.1	0.8	3.9	6.8	15.4	25.1	29.4	16.1	1.9	1026
Thailand	0.0	0.0	1.0	6.7	16.8	23.5	27.9	17.8	5.8	0.4	0.0	0.1	937

The distributions suggest that in Germany, the Netherlands, and Sweden in Population II and in Germany, Finland, the Netherlands, and Sweden in Population IV, the majority of students have a knowledge of English which is slightly above what might be expected of them, while those in some other countries tended to achieve a lower level than might be expected of a sample defined according to the criteria of this study.

Speaking and Writing

Table 15

Percentage Frequencies for Writing and Speaking Tests: Global Rating

Country/Population	Writing 0	1	2	3	4	Speaking 0	1	2	3	4
Population II										
Belgium	13.3	36.2	42.1	7.4	1.0	21.6	39.6	34.7	3.5	0.5
Finland	49.1	20.2	20.7	7.8	2.2	38.2	22.8	28.1	8.1	2.8
Italy	27.9	42.6	26.1	2.8	0.6	23.3	63.3	12.5	0.9	0.0
Sweden	19.2	38.9	31.4	9.2	1.3	19.5	40.8	31.0	6.2	2.4
Population IV										
Belgium	6.5	21.8	54.5	15.3	2.0	3.8	35.8	48.3	12.1	0.0
Chile	39.4	42.2	16.9	0.9	0.6	69.5	26.5	4.0	0.0	0.0
Finland	1.3	20.9	57.6	17.7	2.6	5.8	28.0	60.8	4.8	0.7
Hungary	7.4	22.3	39.6	23.0	7.7	2.5	24.8	28.2	26.7	17.7
Italy	6.1	35.7	45.8	12.2	0.2	0.0	42.3	40.2	17.5	0.0
Sweden	1.9	17.4	43.3	32.0	5.4	3.8	23.6	56.9	13.1	2.6

Speaking and Writing tests were given to only a small sample of students in both Populations. Distribution frequencies were prepared for each of the sections of the 2 tests, but not the Total Speaking or Total Writing Scores. Global Ratings are the only sectional scores provided for both Speaking and Writing. The criteria of definition for all the other sectional scores differ. However, the Global Rating Scores, in spite of precautions to ensure inter-rater stability of scoring, reflect the subjective judgements of the scorers. The Writing test distribution for Population II suggests that the level of general writing proficiency was regarded by the scorers as lower than they expected, while in Population IV the scores approximated to a normal distribution curve. The same is true of the Speaking Test Scores.

ANCHOR ITEMS

Sixteen items of the Reading test and 12 of the Listening test were common to both Populations. Table 16 supplies the Means and Standard Deviations of the Reading and Listening Anchor Items, together with the same information in respect to the Total Reading and Listening Scores. Information is thus available concerning the behaviour of the Anchor Items compared with that of the other items in the 2 tests.

Table 16

Means and Standard Deviations for Total Scores and Anchor Items: Reading and Listening

	Reading				*Listening*			
	Total Scores		Anchor Items		Total Scores		Anchor Items	
Country/Population	M	S.D.	M	S.D.	M	S.D.	M	S.D.
Population II								
Belgium	22.8	15.2	6.1	2.7	9.4	5.7	7.5	2.2
FRG	29.9	16.2	7.6	2.9	-	-	-	-

	Reading				Listening			
	Total Scores		Anchor Items		Total Scores		Anchor Items	
Country/Population	M	S.D.	M	S.D.	M	S.D.	M	S.D.
Finland	16.7	17.8	5.9	2.8	8.1	6.2	7.3	2.2
Israel	26.3	18.7	7.0	3.2	-	-	-	-
Italy	18.6	17.3	5.2	3.1	8.9	6.4	7.3	2.2
Netherlands	37.4	16.4	8.2	2.6	-	-	-	-
Sweden	35.0	14.9	8.4	2.6	15.5	4.6	9.3	1.6
Thailand	18.4	13.4	5.9	2.3	-	-	-	-
Population IV								
Belgium	24.2	10.8	9.0	2.2	16.9	7.0	8.6	1.9
Chile	10.6	9.8	5.8	2.5	7.1	4.5	5.5	2.2
FRG	41.2	6.1	11.3	0.8	-	-	-	-
Finland	35.6	8.3	10.9	1.2	23.2	6.1	9.6	1.6
Hungary	20.1	11.2	8.6	2.2	16.3	7.3	8.1	2.0
Israel	27.4	12.9	9.5	2.3	-	-	-	-
Italy	20.2	12.6	8.1	2.6	14.7	7.6	7.7	1.9
Netherlands	42.9	6.5	11.3	0.9	-	-	-	-
Sweden	39.5	9.0	11.0	1.3	26.2	5.4	10.0	1.3
Thailand	19.7	9.1	8.9	2.0	-	-	-	-
No. of Items	60		16		36		12	

INTERCORRELATIONS

Two types of measures, generally referred to as "cognitive" and "affective," were employed in testing the samples studied. The affective measures included scales of interest in English, personal assessment by the student of the utility of English, and a measure of out-of-school participation in English-speaking activities. It has always been understood that achievement in English, as in other subjects, is influenced by the operation of such affective factors. It is for this reason that these possible relationships, their extent and nature, are explored at different levels and in different countries in the present study. Intercorrelations were computed for Populations tested in each country, and the tables take only a selection of variables into account.

Tables 17-20 present the intercorrelations for the cognitive and affective measures for only Belgium and Sweden. They appear to be the most representative, and lack of space makes a choice necessary. Caution should be exercised in interpreting these correlation coefficients, since the reliabilities of some measures were not computed (e.g., Speaking, Writing 1, 2, 3; Writing 4). Like the reliabilities which have been computed, these are likely to vary considerably. Further, some correlations are based on data obtained from a subsample (e.g., Speaking and Writing 4) which may or may not have an influence on the size of the correlation coefficients.

Inconsistency from country to country, as well as between Populations, tends to be the main characteristic, even when corrections for attenuation are made where reliabilities have been computed. Such between-country and between-Population differences may arise from the nature of the abilities which were tested and from the kinds of tests which were used, as well as from the skills, attitudes, and motivations of the students being tested on these particular tests. Differences may also be attributable to factors which have to do with types of courses and curricula followed by the students. Against this background of between-country and between-Population differences, there is a perceptible tendency for the affective measures

Table 17

Intercorrelations for Cognitive and Affective Measures and Number Responding to Each Measure: Population II, Belgium (Fr.)

Each correlation represents the relationship between 2 measures on the same individuals. Since the number responding to each measure varies, the entire sample was not always used to compute the correlations. However, for those measures showing a marked decrease in number responding as compared to the Reading test, the sample represented is generally the result of the best possible subsample of the total sample per country.

	1	2	3	4	5	6	7	8	No.* Responding to Each Measure
1. English Reading	-	.66	.25	.20	.28	.49	.54	.84	687
2. English Listening		-	.25	.16	.24	.51	.37	.65	670
3. Perceived Utility of English			-	.28	.52	.15	.20	.23	677
4. English Activities Outside School				-	.26	.22	.17	.19	678
5. Interest in English					-	.15	.18	.22	678
6. Total Speaking						-	.34	.49	162
7. Total Writing 4							-	.51	680
8. Total Writing 1, 2, 3								-	696

*This is the weighted number used in computing the correlations.

Table 18

Intercorrelations for Cognitive and Affective Measures and Number Responding to Each Measure: Population IV, Belgium (Fr.)

	1	2	3	4	5	6	7	8
1. English Reading	-	.64	.29	.28	.27	.62	.50	.75
2. English Listening		-	.24	.28	.23	.63	.40	.62
3. Perceived Utility of English			-	.36	.59	.20	.23	.26
4. English Activities Outside School				-	.37	.26	.23	.26
5. Interest in English					-	.27	.25	.25
6. Total Speaking						-	.54	.59
7. Total Writing 4							-	.52
8. Total Writing 1, 2, 3								-

Table 19

Intercorrelations for Cognitive and Affective Measures and Number Responding to Each Measure: Population II, Sweden

	1	2	3	4	5	6	7	8	No. Responding to Each Measure
1. English Reading	-	.80	.25	.20	.49	.69	.57	.81	2331
2. English Listening		-	.21	.21	.45	.66	.55	.78	2277
3. Perceived Utility of English			-	.35	.52	.31	.21	.22	2168
4. English Activities Outside School				-	.29	.24	.21	.21	2168
5. Interest in English					-	.49	.36	.46	2171
6. Total Speaking						-	.44	.66	218
7. Total Writing 4							-	.61	2009
8. Total Writing 1, 2, 3								-	2279

Table 20

Intercorrelations for Cognitive and Affective Measures and Number Responding to Each Measure: Population IV, Sweden

	1	2	3	4	5	6	7	8	No. Responding to Each Measure
1. English Reading	-	.62	.19	.29	.18	.43	.41	.64	1619
2. English Listening		-	.12	.28	.14	,41	.34	.51	1588
3. Perceived Utility of English			-	.38	.55	.27	.21	.16	1436
4. English Activities Outside School				-	.37	.41	.33	.27	1447
5. Interest in English					-	.26	.16	.15	1447
6. Total Speaking						-	.24	.30	204
7. Total Writing 4							-	.47	1489
8. Total Writing 1, 2, 3								-	1526

to have low correlations among themselves. The cognitive measures tend to have moderate to high positive correlations among themselves. In Hungary, Finland, Italy, and Sweden, there are some cases where the correlations between certain cognitive and certain affective measures, on the one hand, are higher than the correlations among some of the cognitive measures themselves. Interest in English may have a higher correlation with achievement in Speaking than the latter with achievement in Listening. Similarly, in Sweden at the Population IV level, English Activities Outside the School may be more highly correlated with achievement in Speaking than is the latter with achievement in Writing Essays.

ITEM ANALYSIS

Tables 21 and 22 provide "item analysis" tabulations for the items which have been described earlier in this article. For each item analysis, the following results are available, but in the tables presented here, the "point biserial correlations" and some other information are omitted.

- Sum of Weights: this is the simple sum of student weights for students who took the instrument;
- Mean: this is the weighted arithmetic mean of scores;
- Standard Deviation: this is the weighted standard deviation of scores;
- Reliabilities: (a) KR20--Kuder Richardson Formula 20 and (b) Horst--Horst Reliability Formula. This is the adjusted reliability which takes into account the variability in item difficulties. It is considered a more valid estimate of reliability. (Note: the range of the reliability co-efficient computed on the basis of internal consistency procedures can range from +1.0 to minus infinity. In some cases, negative reliabilities are presented. They have not been corrected to 0.)

For each item, the following are reported:

- The item number;
- The correct answer, expressed in numerical form (1=A, 2=B, etc.);
- The weighted item difficulty: the sum of the weights for individuals who answered the item correctly, divided by the total sum of weights;
- Point Biserial: weighted point biserial correlation between the item and the score based on the remaining items;
- Percent Response Distribution: for each response, the sum of the weights for individuals choosing that response divided by the total sum of weights; this is also done for those individuals who did not reach the item (NR);
- Point Biserial Correlations: weighted point biserial correlations between each response and the Total Score except for the correct response, where it is an item-remainder correlation. (Note: the Stage 2 analyses did not produce item-remainder correlations in the item analyses, only item-total correlations which are spuriously inflated, especially in short sub-tests.)

Analysis of Data for Listening Test Items Referred to in this Article, by Country and Population

Country	Sum of Wts.	% Diff.	Point-Bis.	Omit	A	B	C	D	E	N.R.	Sum of Wts.	% Diff.	Point-Bis.	Omit	A	B	C	D	E	N.R.
	Item 13 (C=correct response)										**Item 19 (A=correct response)**									
Population II																				
Belgium	670	.65	.27	1.1	7.7	26.0	65.2	-	-		670	.69	.22	2.9	69.4	12.6	15.3	-	-	
FRG	6	0.0	0.0	0.0	100.0	0.0	0.0	-	-		6	0.0	0.0	0.0	0.0	0.0	100,0	-	-	
Finland	2118	.77	.34	0.2	11.7	11.6	76.5	-	-		2118	.73	.35	.6	73.2	14.3	11.8		-	
Israel	17	.12	.05	16.5	6.9	35.5	12.4	-	-	16.5	17	.14	.57	13.2	13.5	0.0	18.9		-	54.0
Italy	694	.73	.24	5.0	5.2	16.4	73.3	-	-	C.1	694	.58	.51	8.0	58.2	14.8	18.1	-	-	1.0
Netherlands	5	.5	-0.6	0.0	0.0	50.0	50.0	-	-	0.0	5	.38	.77	50.2	37.6	0.0	0.0	-	-	12.2
Sweden	2268	.9	.32	.1	3.9	4.5	91.5	-	-	0.0	2268	.93	.25	.1	93.0	2.1	4.7	-	-	0.0
Thailand	12	0.0	0.0	22.6	35.7	22.6	0.0	-	-	19.0	12	.18	.70	27.4	17.8	17.8	0.0	-	-	37.0
	Item 13 (D=correct response)										**Item 14 (C=correct response)**									
Population IV																				
Belgium	1394	.85	.16	.6	3.5	4.7	3.7	85.0	2.4	.1		.44	.19	2.3	6.3	7.6	43.6	16.3	23.7	.1
Chile	231	.71	.39	4.3	4.7	6.0	7.2	71.1	4.1	2.5		.12	.08	15.7	14.0	14.6	11.7	24.3	17.1	2.5
Finland	2296	.8	.23	0.0	4.6	8.1	6.4	79.6	1.3	0.0	2296	.5	.09	0.6	7.2	13.7	50.0	14.5	14.1	0.0
Hungary	1062	.9	.2	.1	3.1	4.7	4.3	86.3	1.6	0.0	1062	.49	.08	.1	6.2	13.4	48.7	19.9	11.7	0.0
Israel	10	0.0	0.0	31.2	28.2	22.2	0.0	0.0	14.3	4.1	10	.1	.35	43.2	4.1	28.2	10.2	6.2	4.1	4.1
Italy	314	.8	.37	.5	.6	2.5	11.8	79.3	.3	5.0	314	.2	.16	6.2	2.1	11.7	22.4	34.3	18.3	5.0
Netherlands	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Sweden	1601	.9	.4	.3	1.8	2.5	2.8	88.0	3.2	1.4	1601	.33	.1	.9	6.3	10.3	32.8	15.0	33.4	1.4
Thailand	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
	Item 21 (B=correct response)										**Item 35 (A=correct response)**									
Population IV																				
Belgium	1394	.46	.22	8.6	14.3	46.2	30.7	0.0	0.0	.2	1394	.56	.36	.7	55.6	15.7	24.6	0.0	0.0	3.4
Chile	231	.36	.11	15.6	9.8	35.8	35.5	0.0	0.0	3.4	231	.29	.19	3.2	29.7	22.2	39.7	0.0	0.0	3.2
FRG	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Finland	2296	.45	.23	.7	11.8	44.9	42.6	0.0	0.0	0.0	2296	.93	.23	0.0	93.5	3.2	3.3	0.0	0.0	0.0
Hungary	1062	.46	.18	1.8	17.4	45.6	35.3	0.0	0.0	0.0	1062	.59	.4	.6	58.6	15.9	24.8	0.0	0.0	.1
Israel	10	.04	.2	28.9	32.5	4.1	24.1	0.0	0.0	10.4	10	.12	.2	12.0	12.2	24.5	16.0	0.0	0.0	35.3
Italy	314	.4	.38	11.7	16.2	41.0	25.6	0.0	0.0	5.5	314	.56	.4	0.0	55.6	9.0	27.0	0.0	0.0	8.5

Table 21, continued

Netherlands	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0
Sweden	1601	.64	.3	1.1	9.2	64.1	24.2	0.0	0.0	1.4		.95	.46	0.0	95.0	2.2	1.6	0.0	0.0	1.5
Thailand	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100.0

Table 22

Analysis of Data for Reading Test Items Referred to in this Article, by Country and Population

Country	Sum of Wts.	% Diff.	Point-Bis.	Omit	A	B	C	D	E	N.R.	% Diff.	Point-Bis.	Omit	A	B	C	D	E	N.R.
	Item 1 (C=correct response)										Item 12 (A=correct response)								
Population II																			
Belgium	687	.47	.34	10.0	7.1	12.1	46.8	8.3	15.7	0.0	.43	.25	1.5	42.6	43.0	4.2	1.3	7.4	0.0
FRG	1074	.8	.2	1.3	1.0	1.9	83.7	3.3	8.8	0.0	.3	.3	3.5	26.3	4.2	5.1	19.6	41.1	.2
Finland	2100	.6	.5	1.7	9.3	9.9	58.4	12.4	8.3	0.0	.15	.13	1.2	15.1	7.0	8.8	27.8	40.2	0.0
Israel	1065	.8	.4	1.9	1.9	4.5	82.8	5.4	3.5	0.0	.1	.23	2.8	14.0	5.3	2.6	20.1	54.9	.3
Italy	791	.6	.4	6.5	3.7	10.8	63.6	4.0	11.4	0.0	.3	.5	3.0	33.7	2.4	3.7	12.6	44.5	.2
Netherlands	2090	.7	.5	.9	2.7	3.9	71.6	7.1	13.8	0.0	.3	.3	.9	31.7	5.0	6.3	25.3	30.8	0.0
Sweden	2322	.8	.5	.2	4.6	3.2	80.5	5.0	6.5	0.0	.4	.4	.4	42.8	8.2	10.1	11.7	26.8	.1
Thailand	1951	.8	.4	.1	2.7	4.0	80.8	5.2	7.2	0.0	.1	-0.1	.5	11.8	6.8	12.3	33.5	35.2	0.0
	Item 6 (E=correct response)										Item 12 (B=correct response)								
Population IV																			
Belgium	1440	.29	.13	2.4	19.6	21.6	18.1	9.6	28.6	.1	.16	.14	14.7	2.4	16.2	47.5	7.9	11.2	.1
Chile	2117	.29	.18	10.4	18.9	14.7	16.8	9.6	29.2	.4	.16	.17	24.0	6.8	16.1	32.2	7.6	12.5	.8
FRG	1374	.77	.19	.2	8.9	8.4	4.2	1.2	77.0	0.0	.25	.18	1.7	1.6	24.9	59.8	9.6	2.4	0.0
Finland	2310	.48	.22	.6	13.7	15.3	14.7	8.1	47.5	0.0	.53	.36	1.6	2.7	52.6	26.7	2.9	13.5	0.0
Hungary	1062	.37	.07	1.7	17.4	19.3	13.7	10.4	37.4	0.0	.3	.06	8.1	2.6	30.6	45.4	5.0	8.2	0.0
Israel	604	.47	.27	3.5	11.8	16.1	12.0	9.3	47.0	.3	.5	.3	4.6	3.2	50.7	13.5	3.6	23.5	1.0
Italy	324	.37	.13	.5	18.6	20.0	18.5	4.5	36.0	0.0	.29	.17	16.0	1.7	28.7	42.4	6.5	4.7	0.0
Netherlands	1560	.79	.28	.1	3.5	10.6	4.6	1.8	79.0	0.0	.47	.25	1.3	2.0	46.8	32.3	3.3	14.3	0.0
Sweden	1626	.68	.4	.3	7.9	11.9	10.5	4.2	65.2	0.0	.9	.17	.2	2.3	91.5	2.6	2.9	.5	0.0
Thailand	937	.24	.01	1.1	16.0	30.7	13.3	14.9	24.0	0.0	.58	.12	.8	4.4	58.5	29.2	2.8	4.4	0.0

ITEM ANALYSIS AND DIAGNOSIS OF ERROR

Failure in Test Design

Errors arose from several sources, one of which was a degree of failure in the preparation of the tests themselves. This failure may be due to a formal or design weakness, where, for instance, one of the distractors, though strictly speaking not a correct answer, does not sufficiently differentiate itself from the correct response to constitute a fair test to the average student. In Population II, *Item 39* of the Reading test read:

(Item 39)

Jack said he did want to come.

A I don't think I shall come.
B I really would like to come.
C I think I shall come.

The correct response is *B*, which takes account of the emphatic nature of the original statement--"did want." But quite considerable numbers in many countries selected the more neutral statement *C* which, though correct so far as it goes, does not convey the intended meaning. This failure in design is equally apparent in *Item 33* of the Population IV Reading test.

(Item 33)

I thought I had seen the car and the girl driving it yesterday.

A I thought I had seen the car and its driver before.
B I saw the car before I saw the girl.
C I thought I had seen a girl driving a car.

The correct response is *A*, which conveys the definiteness of the original statement--"the car" and "the girl," rather than the more open and indefinite statement *C*, which was selected by large numbers of students. Except in Sweden, where the proportion of students selecting *C* was 33% compared with 65% for the correct response, the percentage choosing the distractor was over 50% in the Netherlands and nearly 75% in Belgium, Hungary, and Thailand. The distractor is too close to the designated response, so that it constitutes an unfair test to the average or even slightly above-average student.

The overall design of the test, though not affecting the performance of students unduly, may have favoured some students. This happened where similar test items were placed in close proximity, this being one of the disadvantages of designing the complete test in homogeneous sections. In these cases there may have been a "practice effect," since the performance on the second of the 2 closely related items appears to have been boosted. For instance, the following pairs tested the same aspect of English in each case--Listening, Population II, *Items 13* and *15, 14* and *18, 16* and *20,* and *17* and *19.* The scores on the second of the items in each pair are almost always higher than the scores on the first, in all countries.

These are formal or design features which may have tended to distort the scores on particular items, but there are also substantive characteristics which may have tended to do so. This is particularly the case where the selection of items or of certain features of individual items may have proved more difficult than had been suggested by the results of pretesting. Reference has been made already to *Items 1-8* of the Population IV Reading test, where the students were asked not so much to produce or reproduce correct stress patterns but to analyze them and to institute comparisons. The complexity of this exercise proved too much for otherwise good students, and the scores in all countries are lower by far on this set of items than they are in other sections.

The same is true of Section II of Population IV Reading, *Items 9-16*, where the students were required to match an adjective to one of 5 nouns. In spite of the favourable results of the pretest, this section proved exceedingly difficult, and the items were not attempted by higher percentages than was the case with other items or sections. The undoubted difficulty of these items arose from their excessive sophistication.

Native Language Interference

Other errors are due neither to failure in the design of the test nor to the inherent difficulties which the English language poses to all students alike. These difficulties are imposed upon the students by the characteristics of the languages the students already know, their native languages. It is apparent from the analysis of some items that this tendency exists, and that some countries have found some items more difficult and others less difficult because the national language has either tended to handicap or facilitate the acquisition of those aspects of English. This may be so for students in Chile, where *Items 17-24* of the Population IV Reading test are concerned, since it is evident that students in that country have a greater problem in handling the infinitive construction than students in other countries, and this may arise from the fact that Spanish does not use this construction to the extent, or in the way, that English does. However, it is not a difficulty the Chilean students share with speakers of other Romance languages, where the same lack of congruity with English exists--for instance, those in French Belgium and Italy. A national tendency is also observable in Finland in respect of the use of prepositional adverbs of duration, the basis for *Item 20* in the Population II Reading test.

It is the Listening test, however, that exemplifies this tendency towards possible distortion by the native language most clearly. The focussing of the question by the location of a rising intonation, characteristic of English, is not prevalent in Spanish, and it is significant that Population IV students in Chile found *Item 14* of this test very difficult. The difficulty they also found in distinguishing between "sheep" and "cheap" and "yet" and "jet" may not be unassociated with the influence of Spanish on their approach to the sounds of English.

National Centres will be better able to estimate the validity of the "interference" argument, and its strength in their individual cases, by their own detailed analysis of the incidence of distractors, and especially of the choice of "popular distractors." But from an admittedly inadequate survey across countries belonging to many language groups, markedly different from each other and from English, the evidence for the argument does not appear to be convincing. The same popular distractors tend to be selected by students speaking very different languages far more often than they are differentiated by them. Sometimes not only is a popular distractor not common to speakers of the same language in different countries, but it tends to cut across speakers of very different languages.

The fact that on the present set of tests a positive or negative distortion deriving from the national languages cannot be confirmed does not imply that such an influence may not be substantiated in other tests, since it may well be that, in selecting the present test items, the designers by chance avoided items which clearly disadvantaged or advantaged one country more than another. Another reason may be that the teaching of English in the several countries has been directed to take account of the problem of native language interference and has safeguarded itself very successfully against such an influence. If this is the case, since few of the countries in the present investigation have based their teaching on contrastive analysis techniques, it may be argued that there is little point in extending the practice, since they have succeeded very well without recourse to it.

Student Error

Whatever the justification for these arguments may be, pro and con, there are certainly more powerful causes of student error than native language interference. In the first place, students tend to be misled by irrelevancies. For instance, in the Population II Reading test, *Item 8*, students were required to match "island" with "like," since the initial vowel of the former is included in the latter. But students in all countries tended to match "milk" with "like," possibly because of the irrelevent fact that the 2 words share the 2 consonantal sound /l/ and /k/ and because they are both monosyllables. In this sense the error arises from the student's search for a regularity in the item stimulus and response, though the regularity is not required for a correct answer. This occurs in *Item 10* of the same test, where "whole" is the correct response to "over." High percentages selected "one" as the correct response to "over," presumably because the initial vowel is repeated in its written form and catches the student's eye immediately as the obvious regular feature.

Inattentive reading, or failure to take account of the context of the item, is a powerful factor in producing student error. Thus, in *Item 50* of the Population IV Reading test, an attentive reading of the passage will direct the student to the choice of response *B*--"there are good harbours sheltered from the sea," as the reason for there being so many amateur sailors in England. However, high percentages of students in many countries chose *A*--"there are so many sailing boats," though this is the consequence rather than the cause of the existence of so many sailors. It is even more apparent in *Item 29* in the Reading test for Population II, where the failure to read attentively is a clear cause of failure. Students were asked to select one of 3 alternatives to a question about the meaning of "thirsty." The correct response is *C*, which reads, "I wanted something to drink." Large numbers chose *A*--"I wanted to sleep"--presumably because the stem included a reference to being "thirsty at night."

This cause of failure is related to another, namely the tendency to ignore the semantic content of either stem or response and to select a response which echoed a formal, but semantically inappropriate, feature of the stem. For instance, in the Population II Reading test, *Item 25*, the student is asked to select the correct response "go out of" as the meaning of "He plans to leave the house." But many students selected response *A*--"live in"--possibly because of the similarity of /leave/ and /live/, rather than being guided by the deeper semantic similarity of the stem and correct response. *Item 28* brought out the same tendency, for instead of selecting "bring back" as corresponding to "return the book," many chose "turn the pages," possibly because of the correspondence between /return/ and /turn/. The same inclination to be guided by the surface similarity or formal identity rather than by deeper semantic correspondence is exemplified by an analysis of the responses to *Items 33, 36, 37, 51*, and *53* in Population II. So far as what may be regarded as predominantly linguistic error enters into the analysis, it may be argued that the preference for the repetition of formal features as clues to the correct response rather than correspondence of meaning between stem and response is the most important single factor. It has very significant linguistic implications not unrelated to the distinction currently favoured by some linguists between "deep" and "surface" structures. It should be an interesting area for future development of research into error analysis.

There is one set of causes of student error which is, strictly speaking, not linguistic so much as a manifestation of more deep seated inadequacies, namely those arising out of the operation of intellectual processes which, though they express themselves in language, are not exclusively or primarily linguistic. Beyond the first stages of foreign language learning, it becomes progressively more difficult to measure linguistic competence in isolation from general intellectual ability. Other extra-linguistic capacities, like the ability to draw inferences, to come to correct or permissible conclusions, and to elicit the relevant information, are also involved. Thus, the successful reading of passages for comprehension depends not only on an acquaintance with the foreign language but also on the ability to bring to bear cognitive forces which are basically independent of a particular language, though expressible in it. This is certainly the case with the kinds of questions asked in the concluding items of the Population IV Reading test, where fairly advanced intellectual capacities are required, as well as knowledge of vocabulary and of the structure of the English language. It is most apparent in *Items 50-60*, which are concerned with the understanding of a discourse on attitudes to political thinking. But the same factor runs through the operation of nearly all the test items.

FOOTNOTES

[1] Ian Dunlop replaced M. Morette who, due to ill health, resigned from the Committee very early in the investigation.

[2] Though she is not responsible for the shortcomings of this presentation or the development of the English programme, Dr. Carolyn E. Massad (Educational Testing Service), during her tenure as Spencer Fellow at Stockholm, was deeply involved in the processing and interpreting of statistical data and in preparing the final report.

[3] In line with the decision to develop a uniform international set of tests, the participating countries were asked whether they would prefer that the items should conform to the norms of British or American English. They all chose the British model.

[4] Support for these assumptions may be derived from the fact that there are considerable differences between means of global ratings for an original and a second scoring, whereas the means of the final scores based on the original and the second scoring tend not to differ very much. Also, the rater reliabilities of the final scores are generally higher than those of the global ratings.

The Test of English as a Foreign Language

John W. Oller, Jr. & Bernard Spolsky,*
University of New Mexico

HISTORY OF THE TEST

By 1960, there was evidence of a strong demand for a good method of assessing the English language competence of foreign students applying for admission to universities in the United States. At that point, about 25 percent of the interested institutions were using some kind of English test, 35 percent were relying on reports of interviews conducted by an embassy official or by local professors, 15 percent sought evidence of English competence from the students' school records, and the remainder took the applicant's word for his ability. Since many of the embassies actually used a test (usually the American University Language Center Test), it is probable that some kind of test was being used in more than half the cases. But there was considerable dissatisfaction: a 1962 survey of U.S. colleges and universities found that more than 90% felt the need for a better program of testing the English of foreign students.

These ideas were clearly expressed at a conference sponsored in 1961 by the Center for Applied Linguistics, the Modern Language Association of America, The Institute of International Education, and the National Association of Foreign Student Advisors. On the basis of suggestions made at this meeting, the National Council on the Testing of English as a Foreign Language was founded with representation from 30 interested organizations. Funds were provided by grants from the Ford and the Danforth Foundations, and a staff was appointed to develop a test.

In June, 1963, test planning and test writing committees met, established details of the test, and prepared to begin the writing of test items. The test was appropriately labeled the Test of English as a Foreign Language (TOEFL). In order to achieve machine scorability, and for other obvious institutional reasons, a multiple choice format was decided on. The test would consist of 5 parts.

The present form of the examination has been changed very little and can be described as follows (the examples are all taken from Pike, 1973): the first section is labeled Listening Comprehension and contains 3 sub-parts: the first part consists of either a statement or a question which can be answered by selecting 1 of 4 alternatives. For example, the student might hear,

When did Tom come here?
(A) By taxi.
(B) Yes, he did.
(C) To study history.
(D) Last night.

or he might hear a sentence like,

---

*It was hoped that this chapter would be written by one of the early program directors, but at the last minute this proved impossible. The present authors have served on the Committee of Examiners--Bernard Spolsky from 1967-1970 (the last 2 years as Chairman), and John Oller since 1972. We are grateful to Educational Testing Service for documents, and to Lewis Pike for permission to quote his unpublished report. We acknowledge with thanks many conversations with Douglas Stevenson: his doctoral dissertation will clarify a number of issues we touch on.

John dropped the letter in the mailbox.
(A) John sent the letter.
(B) John opened the letter.
(C) John lost the letter.
(D) John destroyed the letter.

In the first case, the student must select the appropriate continuation, and in the second, he must select the best paraphrase for the sample sentence.

In the second part of the Listening Comprehension section, the examinee hears a conversation. For example, he might hear:

(Man's voice): Hello, Mary. This is Mr. Smith at the office. Is Bill feeling any better today?
(Woman's voice): Oh, yes, Mr. Smith. He's feeling much better now. But the doctor says he'll have to stay in bed until Monday.
(Third voice): Where is Bill now?
(A) At the office.
(B) On his way to work.
(C) Home in bed.
(D) Away on vacation.

Usually, the dialogue-type items require some inference on the part of the student.

In the third part of the Listening Comprehension section, the examinee hears a simulated lecture during which he is allowed to take notes. Important terms and names that occur in the lecture are provided on a separate page for note-taking, which is placed in the hands of the examinee. It is presumed that these notes simulate the kind of things that a lecturer might write on the blackboard. The questions concerning the lecture are similar to the kinds of questions that normally follow a reading comprehension passage.

Section Two of the TOEFL, labeled English Structure, consists of items that are supposedly simulated conversations. Each one is a multiple choice type question like the following:

"John needs a pencil."
"He can use one ________."
(A) of me.
(B) my.
(C) mine.
(D) of mine.

Section Three, Vocabulary, consists of 2 parts. The first is a sentence completion task, and the second is a synonym matching task. For example, the student might read,

A ________ is used to eat with.
(A) plow
(B) fork
(C) hammer
(D) needle

or he might be given a stem word and be asked to select the best synonym from a list of alternatives such as,

foolish
(A) clever
(B) mild
(C) silly
(D) frank

Section Four, Reading Comprehension, provides the examinee with paragraphs, each followed by several questions. For instance, the student might read,

The White House, the official home of the President of the United States, was designed by the architect, James Hoban, who is said to have been influenced by the design of a palace in Ireland. The building was begun in 1792 and was first occupied by President and Mrs. John Adams in November, 1800. The house received its present name when it was painted white after being damaged by fire in 1814.

When was the White House first occupied?
(A) 1776
(B) 1792
(C) 1800
(D) 1814

According to the paragraph, the president's house was first painted white when
(A) President and Mrs. Adams requested that it be repainted.
(B) it was repaired following a fire.
(C) the architect suggested the new color.
(D) it was remodeled to look like an Irish palace.

Section Five, the so-called Writing Ability test, consists of 2 parts. The first is an error recognition task, and the second is a sentence completion task. For example, the student might read,

$\underset{(A)}{\underline{\text{At first}}}$, the old woman seemed unwilling $\underset{(B)}{\underline{\text{to accept}}}$ anything $\underset{(C)}{\underline{\text{that}}}$ was offered to her by my friends and $\underset{(D)}{\underline{\text{I}}}$.

The task for the student is to recognize the fact that the item lettered D is an error according to some standard version of a prescriptive grammar for English.

An example of a sentence completion problem is:

Because he had little education, his knowledge of the subject was ________.
(A) limited
(B) small in quantity
(C) minor
(D) not large at all

The first form of the TOEFL was ready for international administration in February, 1964; 920 candidates took it in 34 countries. In July, 1965, responsibility for the test was passed to the College Entrance Examination Board (CEEB) and Educational Testing Service (ETS), with appropriate advice still available from the National Advisory Council. CEEB and ETS have, in principle, equal responsibility: in practice, ETS handles all operational details, and CEEB watches the interests of candidates and institutions that use TOEFL. The program has grown rapidly: 2,482 candidates were tested at 80 centers in 1964-65, but 64,243 were tested at 1,582 centers in 1971-72 (Jameson and Malcolm, 1972). It is now administered at centers in about 120 countries (ETS, 1973).

The structure of the test was established by the first conference, refined by the first program staff and committees, and has since 1965 been under the strong institutional control of ETS and CEEB.

Shortly after ETS and CEEB assumed control, a number of major administrative changes took place. There was a great increase in the number of tests given, and some steps were taken to improve the quality of the test. For instance, a 6-member Committee of Examiners with revolving membership was set up in 1966 to advise on test development and research. From 1963 until 1969, all new questions were first pretested with groups of foreign students at American institutions; after 1970, an "experimental" section of untried items was included in all domestic administrations. A large pool of questions has thus been developed, and a number of research projects of various dimensions were carried out. [We will discuss the most recent and extensive of these studies (Pike, 1973) below.] Since 1970, candidates have received their scores and a booklet on how to interpret them. Administrative functions such as registration of students, administrations of tests and test centers, and reporting of results have been transferred where possible to local agencies: this is now done in Taiwan, Hong Kong, and India. The number of scheduled administrations was raised in 1969 from 3 to 4, and some forms of the test have been made available for colleges and universities to use for their own purposes. A walk-in test center was established in 1970 at the ETS Office in Washington, D.C., and several more have now been set up elsewhere in the United States. To handle difficult problems of test security, especially to forestall the danger that cram school operations in various parts of the world might compromise the test, a new equating system was developed that did not require reference to questions previously used for equating purposes.

All these administrative improvements have not been matched by changes in the form of the test: why this has been the case will be discussed in more detail below. Basically, it is fair to say that in its first decade of operation, TOEFL has become an efficiently-administered and widely-respected test.

ASSUMPTIONS UNDERLYING TOEFL

Several principles were agreed to in 1963 by the Test Planning Committee of the National Council on the Testing of English as a Foreign Language:

- Contrastive analysis as a basis for test construction was rejected. The purpose of the test was not to discover and test "special problems encountered by students from differing language backgrounds" (NCTE, 1963, p. 2). This could be understood both as a practical recognition of the heterogeneity of the people to be tested and as a theoretical rejection of Lado's view of testing: to wit, "the theory assumes that testing control of the problem is testing control of the language" (Lado, 1961, p. 24).
- Vocabulary control was to be exercised wherever possible in order to keep reading comprehension at a low level of difficulty.
- The test items were to be written in a way that would present language "in as realistic a context as possible" (NCTE, 1963, p. 2).

In general, these were the explicit guidelines laid down by the Test Planning Committee. There were some additional assumptions, however, that are implicit in the development of test forms and in most of the research that has been done on the TOEFL ever since the Test Planning Committee first met. The most important of these unstated assumptions derives from psychometrics. For example, there has been a consistent emphasis on the importance of statistically acceptable items and subsections. With this in mind, the test has been carefully standardized in reference to a sizable population of non-native speakers. In keeping with the sound psychometric maxim of knowing what you are measuring, there has been an attempt to develop an "omnibus battery testing the wide range of English proficiency and yielding meaningful (reliable) sub-scores in addition to total scores" (The Planning Committee, as cited by Slocum, 1969, p. 2). The attempt to differentiate skills and also the inclusion of both discrete item sections (English Structure, Vocabulary, and Writing Ability) and more integrative sections (Listening Comprehension and Reading Comprehension) probably reflected the influence of Carroll (1961).

No one who understands the utility of statistical procedures of the sort employed by ETS and CEEB could seriously deny the importance of the emphasis placed on psychometric procedures. However, from a linguistic vantage point (whether psycho- or sociolinguistic, or both), it is just possible that the assumption that the TOEFL should be standardized on the basis of non-native speaker performance, and the emphasis on independence of sub-scores, may actually limit the effectiveness of the test. There is now a considerable body of research which suggests that the subscores on the TOEFL, or any other language proficiency test, are apt to be highly interrelated. In fact, strong arguments have been developed showing that, from a linguistic point of view, low correlations indicate a lack of validity in the sub-parts of a language proficiency test (see especially Irvine et al., 1974).

By the same line of reasoning, the standardization of TOEFL against a population of non-native speakers may be questioned. The problem of evaluating language proficiency can be viewed as one of determining how closely the non-native speaker's skill approximates that of a native speaker. If the test is to be used primarily as a basis for estimating "the readiness, in terms of English proficiency, of non-native English speakers for study in the United States [at the college or university level]" as Slocum (1969) has said, perhaps it would make more sense to standardize the test in reference to the performance of native speakers at that level rather than in reference to a large group of non-natives (*cf*. Darnell, 1968).

If a test were developed strictly on the basis of psychometric procedures without calling into play what is known of language competence from a linguistic point of view, it is just possible that it could become increasingly biased as items were constructed more and more to conform statistically to the unique characteristics of a somewhat unusual population of speakers. This is contrary to the view that the best standard of language proficiency is the native speaker. It may not be possible to say *exactly* what is meant by "the proficiency of the native speaker," but it is possible to say *in general* how well the native speaker is able to perform certain types of language tasks. This performance will differ substantially from that of typical non-native speakers. Native speakers may perform less well in some cases than non-natives, but on language based tasks, they should normally do better than non-natives. It is precisely the kinds of tasks that native speakers perform almost flawlessly that language proficiency tests should probably consist of.

Unfortunately, all sections of TOEFL have not always focussed on such tasks. For example, a study by Angoff and Sharon (1971) revealed that non-natives did better than native speakers of English on more than 21 percent of the items in the Writing Ability section of a certain form of TOEFL. Also, some TOEFL Vocabulary items, though they may yield good statistics, tend to rely on esoteric terms that are of limited usefulness, e.g., "ogle," "chortle," "quash," and the like. Both of these problems may be due to standardization of TOEFL against a population of non-natives and to the concomitant methods of statistically evaluating items in reference to that population.

Another assumption implicit in TOEFL research concerns the correlation between parts of that examination. The *Manual for TOEFL Score Recipients* (ETS, 1973) states that low correlations between Listening Comprehension and other parts of the test can probably be explained "because skill in listening comprehension may be quite independent of skills in reading and writing" (p. 15). This claim is probably related to the original motivations for reporting separate sub-scores on the test. "The test was to consist of five sections, each yielding a subscore sufficiently reliable for diagnostic purposes" (ETS, 1973). However, there is a considerable amount of research that suggests that this conclusion, plausible though it seems, may well be incorrect. For most non-native speakers, the ability to listen with comprehension is highly interrelated with reading and writing skills. This is best demonstrated in some of the research with cloze procedure, dictation, and other measures of listening comprehension, reading, and writing skills (see articles in this volume and their references). It may be that instead of indicating an independence of skills, low correlations between sub-scores on the TOEFL indicate weak validity in one or more parts of the test.

None of this, of course, is intended to fault the original planning and development of TOEFL. When TOEFL was first being conceptualized, most of the research which now raises questions concerning original assumptions was not yet available. Happily, some of the objections that have been raised recently may be answered by some of the new directions that research on the TOEFL has suggested. It is to be hoped that test development will become even more sensitive to research and that future forms of TOEFL will incorporate its findings.

THE USE AND INTERPRETATION OF TOEFL

In an explanation of the growth and success of the testing program, the 1973 *Manual* mentions 3 factors: the security of the test, its availability throughout the world, and "the high reliability of TOEFL scores for admission and placement purposes." There has been little debate on the first 2 of these, but there has been constant questioning of the third, usually resolved in favor of the test.

One basic issue was discussed in some detail in Spolsky (1967). To what extent is English proficiency (however measured) a useful predictor of the success of a foreign student in a U.S. university? This is an important question independent of its relation to the matter of validating proficiency tests. Unfortunately, several studies have attempted to relate Grade Point Average (GPA) rather more directly to TOEFL scores.

Pike (1973), however, emphasized early in his report that GPA is clearly not a good criterion against which to test the validity of TOEFL, or any language proficiency test for that matter. There are many reasons for this which have been discussed in several previous publications. In fact, the irrelevance of GPA as a validating criterion was noted by Slocum (1969) and reiterated in the *Manual for TOEFL Score Recipients* (ETS, 1973). Nevertheless, some scholars have persistently claimed that TOEFL should predict GPA. Rand (1972) went so far as to argue that GPA is actually the most important criterion for evaluating the validity of TOEFL. But according to Kaplan (1969), TOEFL is only supposed to measure English language proficiency, and there are clearly other factors which influence GPA and which are not measured by any test of language proficiency. According to Kaplan, those other factors must include "aptitude, attitudes, and motivation" (p. 7).

Unfortunately, publications concerning the TOEFL in relation to GPA have not always taken such a reasonable stand. The booklet entitled *TOEFL Interpretive Information* (ETS, revised January, 1968) included reference to several studies purporting to relate the TOEFL to GPA as a validating criterion. That manual reported correlations between the TOEFL and GPA in studies at the University of California at Berkeley, Massachusetts Institute of Technology, and Fresno State College.

The *Manual for TOEFL Score Recipients* (ETS, 1973) later pointed out that such studies,

> have generally yielded positive correlations between TOEFL scores and grade-point averages, but usually so low as to be of little practical usefulness in the admission process. This result is to be expected if TOEFL scores are used properly. If only those foreign candidates who have demonstrated a high level of proficiency in English are admitted, or if provisions are made through intensive English courses and accompanying modifications in course load to upgrade the proficiency of the less competent, then one would expect that English fluency would not be a major factor in determining academic success. Moreover, the proficiency in English of a foreign candidate--at least at the level measured by TOEFL--is not as stable a characteristic as verbal or mathematical aptitude. Rather it is subject to considerable change over relatively short periods of time. Consequently, a student who may be handicapped because of language problems during his first term of study may not be handicapped after that time. (p. 27)

Slocum (1969) urged that care should be exercised not to interpret scores absolutely:

> One must be very wary of thinking of a TOEFL score as a magic number. It is necessary, I think, to imagine a TOEFL score as being within a certain range of proficiency, and as indicating an ability to perform a certain level of linguistic tasks. Doing so will help one avoid the unfortunate mistake of speaking of "cut-off" scores. (p. 4)

In fact, the several *Interpretive Manuals* that have been produced were at least partly intended to forestall the understandable tendency of admissions officers and other users to look for absolute interpretations of TOEFL scores. In the January, 1967 booklet, normative information is given based on all candidates who were tested in regular administrations from February, 1964 until April, 1966. This information was given in the form of frequency distributions of part scores and total scores. In addition, breakdowns were given for various classifications of students: undergraduates, graduates, major fields, sex, degree plans, nationality, and native language. Reports were also given of a number of research studies involving TOEFL: a comparative study of TOEFL with the American Language Institute Test of New York University and the Michigan Test, and a number of other correlations.

The January, 1968 bulletin added data on candidates through April, 1967 and gave a suggested interpretation of test scores based on experience at some universities. The bulletin encouraged institutions to conduct their own studies but at the same time suggested: (1) no restrictions on students with scores of 550 and above, (2) students between 450 and 549 needed some English as a Foreign Language training but no restrictions in course load, (3) students between 300 and 449 should have reduced loads and considerable English as a Foreign Language, and (4) students between 200 and 299 would probably require full-time intensive instruction in English as a Foreign Language before taking regular subject matter courses.

The 1973 *Manual*, which is the most recently published manual as of this writing, included normative data based on candidates who took TOEFL from October, 1966 through June, 1971. There are breakdowns for part scores, major fields, native language, continent, and native country. When it comes to making specific recommendations for the use of TOEFL scores, this revision of the manual is cautious. A number of decisions, it says,

> require the admissions officer to judge whether the applicant is sufficiently fluent in English to meet the demands of a regular or modified program of study at his institution. Such judgments should not be based on TOEFL scores alone. Inasmuch as TOEFL scores simply indicate the relative positions of candidates from low to high, there is no way of inferring from the scores themselves whether the candidates have the English language skill necessary for successful performance in particular situations at particular institutions. The relationship between TOEFL scores and subsequent performance can only be ascertained empirically, preferably at the institutional level. (p. 26)

Realizing, however, that many institutions are too small to gather their own data, some guidelines are given. First, a report is given on a 1969 survey in which a good number of institutions reported on the lowest scores that they would accept for foreign students. Secondly, the admission policies of two major, but unnamed, universities are given; these policies are modified versions of the interpretation in the 1968 bulletin. Thirdly, some suggestions are made about relating TOEFL scores to other admission tests. As we noted above, due caution was included concerning TOEFL scores in relation to GPA. Finally, the bulletin sets out 5 basic principles which are worth quoting:

> 1. *Consider part scores as well as total scores in making admission decisions.*
> 2. *In arriving at a decision concerning the competency in English of a foreign applicant, give consideration to the different kinds and levels of proficiency required in different fields and levels of study and to the resources available at the institution for developing the English language skills of foreign applicants.*
> 3. *Do not use "cut-off" scores in screening foreign applicants.*
> 4. *Base admission decisions on all available relevant information--not solely on TOEFL scores.*
> 5. *Assemble information on the validity of admission decisions based on TOEFL scores.* (p. 28-29)

FUTURE DIRECTIONS FOR TOEFL

As Kaplan (1969) observed, one of the significant problems which thwarts possible changes in any existing institutional program like TOEFL is the amount of inertia created by the large sums of money already expended. As he puts it,

> . . . by the time a test reaches the present proportions of TOEFL, the administering agency has so much invested that it cannot afford to give up the test in favor of a new direction, and so research and improvement fall victim to economic necessity.

Since the format was set in June, 1963, changes have generally been minor ones and have been made only after long, careful, and sometimes arduous discussions. As a result, the current form of the test tends to reflect the best thinking of the early 1960s.

Although the TOEFL has remained in its present format for approximately 11 years, it now seems fairly likely that substantial changes will be forthcoming in the next 3 to 5 years. Jameson and Malcolm (1972) noted the possibility of incorporating new research findings into future forms. Moreover, the present Committee of Examiners has strongly urged that means for future revision and constant updating be incorporated in the test production and evaluation process. Most of the changes that are presently under consideration follow a study by Lewis W. Pike (1973), although many of them were strongly recommended by earlier Committees of Examiners.

In the initial stages, the Pike study was sometimes referred to as "a validation study." Certainly, it was concerned with the validity of TOEFL, but not in the sense of validating so much as of testing the degree of validity of TOEFL. Later, the title evolved to "An Evaluation of Present and Alternative Item Formats for Use in the TOEFL." The latter is more appropriate, partly because of the difficulty of saying exactly what some of the sub-sections are measuring. About all that one can say with any great confidence is that several sections of TOEFL are providing very useful information concerning global language proficiency, and a couple of the sections are providing somewhat less information on this global proficiency factor. Because of the uniqueness of the Pike study, and its likely importance to the future of TOEFL, it may be useful here to give a very brief summary of the 114-page report circulated in a tentative format in October, 1973.

THE PIKE STUDY

The purpose of the study was "to obtain information useful for evaluating and possibly revising the specifications and content of TOEFL" (Pike, 1973, p. 1). It examined 3 types of problems: (1) How valid is the current form? (2) What should TOEFL measure? (3) What sub-tests should TOEFL contain?

For all of the reasons noted earlier, it is gratifying to see the Pike study concentrating on criteria other than GPA. He used 4 experimental multiple choice tests, 2 objective open-ended tests, and 2 subjectively evaluated measures. The first of the 4 experimental tests is a written version of Part A of the Listening Comprehension section of TOEFL. (See the description of each part of TOEFL above.) The second experimental test required that the student replace an underlined word or phrase with 1 of 4 alternatives. The third was a sentence embedding task which required the examinee to select the best paraphrase of several short simple sentences. The fourth, essentially a cloze task, was converted into a multiple choice test by listing 4 alternatives for each blank. One of the alternatives was the word originally deleted from the blank. The 2 open-ended procedures consisted of a traditional open-ended cloze task and a sentence combining task of the sort reported by Kellogg Hunt (1970). For the latter, the students read several short simple sentences and were asked to rewrite them in a more acceptable format using more complicated structures. In addition, there were two experimental subjective measures. Each candidate took an oral interview test, and each wrote 4 short essays on 4 topics. The interviews were evaluated by 10 listeners, and the essays were evaluated by 2 or more readers. A more comprehensive description of the evaluation procedures, as well as the tests themselves, is given in Pike (1973).

Quite naturally, the question of the validity of the current form of TOEFL is related to the problem of what it ought to be measuring, and so, what sub-tests it should contain. Another way to get at this is to make some assumptions about what each sub-test should be measuring and to see how well the sub-test does what it is supposed to do. Looked at from this point of view, we might expect the Listening

Comprehension section to correlate better with the interview than with any of the other experimental tasks described above. However, when we examine the correlation tables produced in Pike's study, we discover that the Listening Comprehension section in fact correlated better on the whole with the essay measure than with the interview measure. Similarly, the writing ability section correlated about as well with the interview and with the cloze procedure as it did with the essay section of the test. About all that can be concluded from an examination of the correlation matrices is that all of the tasks seem to be highly interrelated. There seems to be little or no justification for the claim that any one section of the test is measuring any one skill any more than any other section. One fairly firm conclusion that can be drawn from the matrices is that the Kellogg Hunt passage did not correlate too well with any of the other measures.

The natural conclusion that Pike, therefore, draws from his study is that TOEFL could be substantially shortened and 1 or more sections could be eliminated. Because of the high correlation between the English structure and writing ability tests, and because of their similarities in other respects, Pike has suggested that they be condensed into 1 section in 2 parts. He has recommended that the Listening Comprehension section be retained in roughly the same format, although he urges that the lecture section be changed to shorter passages or mini-lectures, to parallel the Reading Comprehension section of the examination, which he also suggests be retained in roughly its present format. Thus, the present 5-part TOEFL could become a 3-part examination if Pike's recommendations are accepted.

CONCLUSION

TOEFL is important not just for its widespread use but because of its association with a good proportion of U.S. language testers. It has thus been open to closer scrutiny than many other standardized language tests. Its beginning was exciting, the following decade was one in which operational growth and institutional efficiency were stressed, and there are hopeful signs that once again it will become not just the most efficient but also the most effective language test for mass administration.

TOEFL in Recent Years

Paul J. Angelis, Educational Testing Service

In July, 1973, the original arrangement whereby the College Entrance Examination Board (CEEB) and Educational Testing Service (ETS) held joint responsibility for the TOEFL program was changed. Two factors led to the restructuring which took place. First, ETS had always been more concerned with the operational aspects of the program than with policy matters. Secondly, the number of prospective graduate students taking the test had significantly increased. As a result, the two groups which assumed responsibility for the direction of the program were the CEEB and the Graduate Record Examinations Board (GREB), with the operation of the program remaining with ETS. This is the arrangement under which the TOEFL program still operates.

One development which resulted from this realignment was the change of the National Advisory Council on TOEFL to the current TOEFL Policy Council. This meant a change in both the membership and responsibility of the group in question. Rather than a general group of advisors representing a number of different constituencies, the new Policy Council is composed first of 3 members appointed by the CEEB and 3 members appointed by the GREB. These 6 form the Executive Committee, and an additional 9 members are chosen as representatives of foreign student advisors, admissions officers, government agencies, and teachers of English as a Second Language. Moreover, the group is now charged with making overall policy decisions for the TOEFL program.

The Committee of Examiners is now a standing committee of the Policy Council. While this committee has existed since 1966, the charge of the 6-member group has been expanded to include regular review of test items and overall guidance on item types and test content. A second standing committee, established in 1976, is the Research Committee--a 5-member group which reviews and approves proposals for research and monitors projects as they are being conducted.

Since the early 1970s, the TOEFL program has continued to expand. By 1977, the number of scheduled administrations had risen to 5, and the number of special centers in the U.S., where the test is offered on a monthly basis, had reached approximately 15. For the 1976-77 year, the number of candidates tested had increased to nearly 150,000. Although some of these cases represent individuals who took the test for other reasons, such as professional certification, the vast majority did so as part of their admissions procedures for undergraduate and graduate programs at U.S. colleges and universities.

As a result of the Pike study referred to in the earlier article by Oller and Spolsky, a number of revisions were made in the format of the test. It was the implementation of these revisions which led to what has since been called the "new TOEFL." The principal finding from the Pike study to have an effect on the changes made was the high correlation between the Structure and Writing Ability sections and between the Reading and Vocabulary sections. Since the Structure and Vocabulary sections had also always been the most difficult to interpret in terms of the skills they affected, these 2 were combined with the other 2 sections mentioned. The Listening Comprehension section remained a separate part of the test. Thus, the test is now composed of 3 sections: Listening Comprehension, Structure and Written Expression, and Reading and Vocabulary.

In addition to these new groupings and the accompanying reduction of the test both in terms of the number of items and the time allotted, other changes were made in the types of items used. It will be easiest to consider these changes by section.

In the Listening Comprehension section there are still 3 parts. The first 2 of these are essentially the same as their counterparts in the previous 5-section test:

the only change is the exclusive use of statements in the first part, i.e., questions no longer appear. It is the third part of this section, however, which has been radically changed. Because of problems associated with the previously used simulated lecture, not the least of which were the reliance on memory and the unreality of the presentation, this part was dropped. In its place series of shorter "mini-talks" were instituted. In some cases, these are conversations between 2 or 3 persons. In others, they are monologues. Often the topics center on practical situations such as weather announcements, travel information, or news broadcasts.

Section Two of the test, Structure and Written Expression, contains 2 item types. The first requires the candidate to choose the word or words which will best complete a given sentence containing a blank, with the choices differing syntactically. The second is an error-recognition type similar to one used in the 5-part test. The sentence completion type previously used in the Writing Ability section of the old format has been dropped.

Section Three, Reading and Vocabulary, also contains only 2 parts. In the vocabulary type, all items require choosing the best word to complete a given sentence; there are no synonym items as in the previous TOEFL. The one significant change in Reading is that some practical items are now used such as labels, want-ads, signs, etc.

The equating system used to relate the various forms of the test and to account for any variation in the level of difficulty among the forms remains the same. Therefore, the scores on any form of the newer 3-part test can be interpreted in the same manner as for the older 5-part one. This is not true, however, for the part scores. Since certain sections have been combined, it is not possible to relate reading scores, for example, for those who took earlier tests using the old format with those from the new format.

As far as broader relationships between the two versions of the test are concerned, one factor is important to keep in mind. Listening Comprehension is the only skill which is still tested as a separate section of the test. In the older format, this section was one of 5; now it is one of 3. Since an individual's total score is directly related to his/her score on the separate sections, some shifts may occur. More data from administrations of the new version of the test will be needed before the nature of such shifts, if they exist, can be determined.

The basic assumptions underlying the test remain very similar to those originally stated when the test was designed in 1963. The scaled scoring system which the test uses is still a normative one. The scores a candidate receives indicate his English proficiency in relation to all others who have taken the test. Such a system appears meaningful to those who receive TOEFL score reports, since it is thereby possible to interpret the English proficiency of an applicant for admission in terms of those foreign students who are, or have been, enrolled at a given institution. Surveys indicate that most institutions which receive TOEFL score reports do use part scores in their admissions procedures, with the Listening Comprehension score in particular appearing to be a useful indicator.

The question of how TOEFL scores should be interpreted for purposes of admission to an academic institution remains an important one. The point has always been made that TOEFL scores are not appropriate indicators of future academic success or failure in an academic program. Presumably, admissions decisions would be made first of all on the basis of information regarding an applicant's previous academic achievement. Scores from an English proficiency test such as TOEFL then provide a measure of whether the person in question possesses the English skills required. But the two issues must be related. To say that TOEFL provides no information of a predictive nature is misleading. The key is to determine the role of English proficiency in various programs of academic study. Doing so should help provide practical guides to what TOEFL scores should mean for different students in a variety of programs at the graduate and undergraduate level.

One of the projects planned in the TOEFL research program will attempt to provide such data, as well as to relate TOEFL total and part scores to other direct measures of the 3 skills it attempts to measure--listening, reading, and writing. This project, which began in early 1978, will in essence be a validity study.

In addition to the activities described above, another stage will attempt to provide feedback from foreign students themselves. Learning how students who performed at different levels on TOEFL were actually able to use their English skills during the course of their academic study will go a long way toward establishing the validity of TOEFL as a measure of specialized English skills. Since the research program began in 1976 with the establishment of the TOEFL Research Committee, a number of projects have been initiated. The first of these, investigating the Performance of Native Speakers of English on the Test of English as a Foreign Language, has been completed, and a research report (Clark, 1977) is available from ETS. A second study looked into the comparative performance of non-native speakers on tests such as the Scholastic Aptitude Test and the Graduate Record Examinations aptitude test. This project has also been completed, and a report will be available shortly. A third project, investigating the feasibility of a test of speaking ability as part of TOEFL, is also in progress. Further areas for research will most likely focus on topics such as the use of TOEFL as a measure of gain, determination of which variables are related to performance on the test, and the possible effect of native language and other background on performance in specific test items.

A new version of the *TOEFL Manual for Score Recipients* will appear this year. It will provide data on the 3-part test, as well as a summary of interpretive information. Thus far, data from the 1976-77 administrations indicate that the statistical nature of the test is consistent with the previous version. The reliability figures for the 3 parts of the test range from .863 to .915, with .954 being the reliability for the total test. The standard error of measurement for the 3 sections ranges from 2.46 to 3.21, with the figure for the whole test being 16.06. Inter-correlations among the 3 parts range from a low of .656 between Listening (Section One) and Structure and Written Expression to a high of .815 between Structure and Written Expression (Section Two) and Reading and Vocabulary (Section Three).

In October 1977, an invitational conference was held in which 10 specialists in various aspects of English as a Second Language participated. The purpose of the meeting was to provide greater input from those familiar with this field to the content of the test. As a result of discussions among these ESL specialists, the TOEFL Committee of Examiners, and ETS staff, some possible shifts in emphasis for the test may occur. The principal suggestions made were for the test to use more extended contexts and more realistic situations in the items. Moreover, since most people who take the test do so for purposes of admission to a college or university, it was felt that greater emphasis needed to be put on academically-oriented contexts.

The TOEFL test has changed in recent years. But with increased input from the various committees associated with the program, feedback from a number of ESL specialists through conferences such as the one described above, and commitment to research, significant changes could lie ahead. As research in second language acquisition and language testing point out new directions for investigation, it is only fitting that a test of major importance such as TOEFL should keep pace.

The FSI Oral Interview*

Randall L. Jones, Brigham Young University

> "To test reality we must see it on the tightrope. When the verities become acrobats, we can judge them."
>
> *Oscar Wilde*

Each year agencies of the U.S. Government send thousands of employees to overseas posts in countries whose principal language is not English. Many of these people are assigned to language-essential positions and must therefore meet established standards of language proficiency. In some cases, especially in the State Department, promotion beyond a certain grade depends on the satisfactory fulfillment of a language requirement. In all cases, the U.S. Government is desirous of having personnel in overseas positions--regardless of their duties--who have some proficiency in the language of the host country.

The means of attaining the necessary proficiency can vary--government language training, college or high school language courses, residence in the country or in a bilingual home, etc.--but the method used to evaluate it has to be uniform. It is not sufficient simply to ask the person how well he can speak the language. The length of time spent in studying a language or living in a foreign country is likewise not a reliable predictor of language proficiency. And standardized language tests do not measure directly the examinee's speaking proficiency. For this reason, it is necessary for the agency to have a standard means of testing an employee's ability in the language in order to determine whether or not the minimum level of proficiency has been acquired.

The method of testing oral language proficiency at the Foreign Service Institute's (FSI) School of Language Studies is known as the Oral Interview.[1] In an Oral Interview, the examinee converses with two FSI examiners for a period of 10 to 30 minutes. One of the examiners, who is in most cases also an instructor at the FSI language school, is a native speaker of the language. The second examiner is usually a linguist, i.e., a language supervisor, who may also be a native speaker of the language being tested, but who is in any case able to understand enough of the language to make a judgment concerning the examinee's proficiency.[2] Most of the interview is conducted by the instructor, while the linguist observes and makes occasional notes. The conversation usually begins at an informal level, dealing with somewhat trivial topics. The level of sophistication gradually increases until the point is reached beyond which it is evident the examinee cannot speak the language without great difficulty. At the conclusion of the interview, the examiners evaluate the proficiency of the examinee on the basis of a set of absolute standards and assign a rating. All Oral Interviews are recorded for possible future reference.

It is important at this point to stress that what is outlined here is not necessarily typical of all Oral Interviews given at the FSI or other federal agencies, as there are certain differences in technique among testers and different procedures at each language school. It would also be appropriate to point out here that the name "Oral Interview" is both redundant and misleading--redundant because the word *interview* generally implies "oral," and misleading because this type of language test is

---

*I am grateful for the valuable information given to me by Claudia Wilds and John Quinones. I appreciate the helpful comments of Earl Rickerson, Claudia Wilds, and Earl Stevick.

really more than an interview. The term has, however, become commonplace within the government language community and will be used here.[3] It is capitalized to distinguish it from other types of oral interview language tests.

The use of the Oral Interview in its present form at the FSI dates back to 1956. At that time, the language testing program was still rather new, and an inventory of language skills within the State Department had just begun to be kept. Instruction at the FSI stressed oral communication, but there was no standard method used to test oral proficiency in any direct way. When the requirement was levied to test the language proficiency of Foreign Service Officers prior to assignment, a type of standardized testing procedure had to be decided on. It was felt that if an officer needed to speak the language in his assignment, then his *ability to speak* should be tested, not his understanding of the grammar or his ability to discriminate the sounds. A face-to-face interview was thought to be the type of test that would afford the examinee the best opportunity to exhibit his skill in speaking the language.

The interview itself was to be as natural as possible, employing the use of artificial stimuli only when the examiners felt that circumstances warranted it. They were responsible for "exploring" the examinee's realm of language proficiency, trying to construct a representative profile as quickly as possible. The evaluation of the examinee's performance in the interview was to be based on an absolute scale of proficiency, with points on the scale reflecting varying degrees of language ability. The tasks involved in making the test, therefore, included developing appropriate interview techniques, defining points on the scale, establishing the criteria for making ratings, and--perhaps most important--training testers to administer the test and make accurate and consistent judgments.

Devising testing techniques was simplified somewhat by the fact that the interview was to be as close to a natural conversation as possible. With practice, each testing team developed useful devices to elicit speech from the examinee. The primary task was to design the method for evaluation. The range of proficiency was divided into 6 levels, *0-5*. The lowest level represented no practical proficiency in the language; the highest represented proficiency equivalent to that of an educated native speaker. Brief definitions were written for the 5 primary levels of proficiency, which were later revised and expanded.[4] In order to make finer distinctions at each level, the testers were given the option of assigning a plus or minus (except for levels *5* and *0*, respectively). The result was a 16-point scale. The minuses were later dropped, since it was found that the scale had become so refined that it was really not possible to make so many discriminations.

To assist the testers in making a rating, a "Check List of Performance Factors" was written.[5] The list contained factors which were deemed critical in the proper use of the language. Early versions of the list contained as many as 14 factors, including grammar, vocabulary, pronunciation, intonation, etc. The list was reduced to 8, and finally to 5, as it became obvious that some of the factors in the early lists were redundant. Each factor was the heading of a 6-point scale with polar terms, based on Osgood's (1957) notion of the semantic differential. The original purpose of the check list was simply to provide a guideline for the tester in assigning a rating. A statistical analysis of some of the earlier lists, however, showed that it was possible to assign numerical values to the various points on the scale, so that a rating could be arrived at by adding the values for each category. Many testers, especially after they had gained a good amount of experience, found that they did not need to use this scoring system in order to determine a rating: their intuitive judgment was as accurate as the check list. Today the check list primarily serves as a training aid for new testers and a useful reminder of the critical performance factors for the experienced ones. All testers are asked to mark the sheet, even if they choose not to tabulate the score.

By 1959, the procedures for administering and scoring the Oral Interview had become fairly well established. The test is basically the same today as it was then. It has, of course, seasoned with age, and the testers are more experienced and confident than at first. The FSI Language Testing Unit, set up in 1958, devoted a significant amount of research time between 1958 and 1968 to studying the Oral Interview, especially its reliability. The results of the studies showed a high consistency among the testers in making ratings.[6] Time was also spent in training new testers and discussing the testing program among the members of the staff. Today the users feel confident that the Oral Interview well serves the purpose for which it was designed, although new ideas and techniques are still being studied and experimented with.

The Oral Interview does have its share of critics, most of whom claim that it is too subjective to be a valid and reliable testing instrument.[7] To the extent that the term has any useful meaning, the Oral Interview would indeed have to be classified as a "subjective" test. This merely means that the examinee does not respond by choosing from suggested alternatives, or by giving a response which is obviously either right or wrong. It is important to keep in mind, however, that a subjective language test can also be very reliable, i.e., consistent in its scoring procedure.[8] Rating an Oral Interview is certainly no more inconsistent than judging a diving meet, a gymnastic tournament, or an ice skating competition. If the judges are well trained and sensitive to the criteria that are important in making the evaluation, their reliability will be high.

In addition to reliability, another important consideration in language testing is validity. A direct test such as the Oral Interview approaches the ultimate criterion of language use far better than any other language test can. John Carroll (1973) has pointed out that the validity of a direct speaking test is high enough to outweigh any problems of reliability. In the case of the Oral Interview, validity and reliability do not depend on how it is constructed--indeed, the test really has no defined structure as such--but rather on how skillfully it is conducted and how accurately the performance of the examinee is evaluated. The success or failure of the test is directly related to how well the examiners have been trained and how well they maintain their ability to conduct a thorough test and make a valid assessment.

The very nature of the Oral Interview insures that every test will be unique. At the beginning, the exchange is for the most part social small talk--"How are you today," "Do you like this rainy weather," "How long have you lived in Washington"--but very soon, if it appears that the examinee is up to it, the conversation will become more serious. The examiners attempt to be as systematic as possible in varying the range of topics and the level of linguistic difficulty, so that the examinee can exhibit a true picture of his language ability. Generally, the nature of the subject matter proceeds from the concrete to the abstract, from discussions of background, family, personal interests, and everyday activities to the expression of opinions on various current events. As soon as it appears that the examinee has sufficiently demonstrated his ability (or lack of it) in one area, the examiners skillfully move on to something else. If the examinee proves that he is incapable of conversing at a certain level or on a particular topic, the examiners rescue him by subtly shifting to another subject.

The role of the tester is not just to ask questions, but to stimulate the examinee to speak as openly and freely as possible by any means available. If the level of proficiency permits it, the interview becomes an exchange of opinions and ideas. If the examinee is speaking because he wants to make what he feels is a significant contribution to an interesting discussion, and not because he is merely answering meaningless questions, he is demonstrating his language proficiency as realistically as is possible under the circumstances. In some cases, the examinee simply does not speak unless he is spoken to, and then only with great economy of words. Good testers claim, however, that they can arouse even the most diffident types to open up at least as much as they would to strangers in a non-testing situation.[9]

Because the language in an Oral Interview is randomly generated, the testers must attempt to elicit a good sample of as many key structures as possible. Even the most capable examiner has occasional difficulty drawing out certain grammatical forms during the interview. Furthermore, the Oral Interview does not provide the examinee with a good opportunity to use certain forms in the language, e.g., interrogatives and imperatives. For this reason, many examiners augment the conversation portion of the test with a less natural, but usually more revealing, language situation. The most common exercise of this type is the so-called interpreter situation. One of the examiners assumes the role of a monolingual speaker of English, while the other one speaks only the target language. The examinee serves as the interpreter for the pair. At lower levels, situations such as getting a room in a hotel, asking directions on the street, or ordering a meal in a restaurant are commonly used. At the higher levels, giving a detailed description of a task to be performed or conducting an interview with an imagined political leader have proven to be useful. With a little imagination, an interpreter situation can become a real-life drama for the examinee. Instead of a routine exchange between the two parties, misunderstandings, differences of opinion, unexpected responses, etc. can lend reality to the situation.

The interpreter situation has the advantage of limiting the possibilities for the examinee. If during a French Oral Interview the examiner asks a question which would normally require a response using the subjunctive mood, but the examinee responds without using the subjunctive in his otherwise accurate and legitimate answer, it does not necessarily indicate that he is not able to use it correctly. In an interpreter situation, on the other hand, the examinee has little choice but to use the structures that are given to him. An additional benefit of the interpreter situation is that comprehension is tested more thoroughly. Care must be exercised, however, in the use of this technique so that it is not overdone, i.e., that it does not occupy an inordinate amount of the total time. The entire exchange should not last more than a few minutes.

Another technique which can be used in an Oral Interview is role playing. A situation is described to the examinee, either orally or in writing, and he is told to play a part. For example, the examinee is told that he is to ask for a room at a hotel. The examiner who is playing the other part must keep the action alive by making it realistic and exercising various options, e.g., "What kind of room would you like," "I'm sorry, we only have two rooms left, and they are both doubles," etc. The possibilities are limitless: reporting an accident, talking to a doctor, changing money in a bank, interviewing for a position, getting a car repaired, putting a classified advertisement in a newspaper, etc. The examinee must be told in some detail what the circumstances of the situation are, either orally or in writing. The main problem with role playing is that its success often depends as much on the imagination of the examinee as it does on his language ability. If the task is not adequately defined, he may not be able to think of anything to say.

A technique similar to role playing can be useful in testing the examinee's ability in using the imperative and interrogative forms. He is given a card on which he is instructed to ask certain questions and make certain requests of one of the examiners. For example, "Mr. Schwarz is your assistant. You will be out of town for a few days, and you need to give him some instructions before you leave. Ask him when and where the June meeting will be, and whether Mrs. Hubner plans to attend. Tell him to call Mr. Mittler and assure him that I will send the papers next week." The examiner who plays the part of the assistant can add a great deal to the situation by making it as realistic as possible. An alternative to this approach is simply to ask the examinee to query one of the examiners about certain topics, e.g., background, personal interests, profession, etc.

A technique which can be used to challenge a high-level examinee is to ask him to describe a common object such as a bookcase, a picture, a pencil sharpener, or a design on a shirt or dress. Many speakers of a second language have an extensive vocabulary and good control of the grammar but are still not able to make precise descriptions of an object. It is, of course, assumed that an educated native speaker of the language would have no difficulty in performing the same task.

When the examiners feel that the examinee has had sufficient opportunity to demonstrate his skill (or lack of it) in communicating in the foreign language, he is excused and a rating is determined. The rating is generally arrived at after the examiners have discussed the test. In some cases, the 2 members of the testing team make an independent rating before the discussion, then try to come to an agreement if there is a discrepancy. The linguist or senior tester always has the final responsibility in deciding the ultimate rating.[10]

A well-trained and experienced examiner is usually able to decide on a rating without a great deal of deliberation. The judgment is largely intuitive, but the intuition is acquired only after a long period of conscious awareness about what should be expected of an examinee at each level. The 2 most important criteria are grammar and vocabulary. Fluency and pronunciation are important, but to a lesser extent. If a speaker is fluent but frequently uses the wrong words and incorrect grammar, his ability for expression is certainly limited. If his pronunciation is foreign but does not interfere with communication, it is not considered critical.

Even though such a method of judging an examinee's language proficiency might appear too unrefined, it has provided surprisingly reliable results. Difficulties do exist, however, especially at the higher levels or in the case of the maverick examinee. There is little problem in distinguishing between a *2* and a *2+*, or a *2+* and a *3*, but the division between *4*, *4+*, and *5* is much less clear. Part of the problem lies with the definition of the *5* level. Just how does an "educated native speaker" speak the language? Is he allowed any errors, and if so what kind and to what extent? Occasionally the examinee *is* a native speaker of the language, but one who has not used it actively for several years. The types of mistakes he makes are often not the same as those made by a native speaker of English who has learned the new language as an adult. Another problematic case is the person who has had no formal training in the language, learned it entirely on his own, and speaks with ease and confidence but with numerous errors. An examiner might be taken in by the fluency of delivery, and overlook basic grammatical and lexical problems.

Anyone interested in using the Oral Interview as a test of language proficiency should keep in mind that it was designed for a specific purpose and has certain shortcomings which might render it an unwise choice in some situations. Among these shortcomings are: (1) the test is expensive to administer, especially to large groups; (2) the test must be administered and scored by trained examiners; (3) the face-to-face situation can create nervousness in examinees; (4) the test cannot provide fine discriminations of proficiency among examinees; and (5) the examiners should occasionally be monitored and retrained. These problems are discussed in more detail in the remainder of the article.

The State Department feels strongly enough about the value of the information provided by the Oral Interivew that it is willing to pay the price necessary to make it part of its language program. A high school or college language department, on the other hand, might find it impossible to afford such a luxury. Each examinee must be tested separately by at least 1 (and preferably 2) trained examiners. Of course, it stands to reason that any oral language test in which the examinee engages in free conversation must either be administered or evaluated on an individual basis. If 50 students go to the language laboratory and record a sample of their speech on tape, each of the 50 tapes must be listened to individually. The same amount of time could be spent in administering and rating the same number of Oral Interviews. The cost can be reduced by using only 1 examiner and by reducing the amount of time for each interview to a few minutes. A few minutes of real spoken language is often more revealing than a 1-hour written examination. Most of the information about the examinee's proficiency is learned in the first 5 minutes of a standard FSI Oral Interview. After that, the time is spent mainly in confirming what has already been suspected. Examinees with proficiency at the higher levels do, of course, require a more thorough test.

The second problem mentioned above concerns the inherent limitations of the testing location. A standardized reading or listening comprehension test is easily sent by mail to any location in the world and can be administered with relatively little training. Unfortunately, this is not the case with the Oral Interview. The FSI does frequently send testing specialists to overseas posts to conduct Oral Interviews, but it is a very expensive practice and can be done only on a limited basis. It would be desirable, for example, as part of the personnel screening process, to test all applicants for employment to the State Department who claim proficiency in a language. But it would simply not be possible to send trained examiners to all parts of the United States for this purpose. Testing by telephone has been tried on occasion, but it has not been used on a regular basis. Another possibility being considered is the evaluation of recorded speech samples which have been taped by a personnel officer during a recruiting visit and sent to Washington for scoring. This is, of course, not an Oral Interview and would serve only as a preliminary indication of the individual's proficiency.

The extent of emotional trauma experienced by the examinee in an Oral Interview is often exaggerated and probably very misunderstood. No matter how pleasant and cordial the examiners may be, an Oral Interview is, after all, a test, and in many cases a great deal is at stake. But emotional strain is part of any test, regardless of whether the examinee is facing a human being, a microphone, or a booklet. It seems logical to assume that the presence of a live tester would heighten the tension, but experience has shown that this is not as significant as one might expect. A skilled examiner is able to put most examinees at ease in a relatively short period of time, and in most cases the examinee soon becomes absorbed in the conversation and forgets about the pressure. The greatest amount of anxiety is shown by those examinees whose proficiency lies at the lower end of the scale.

The Oral Interview was designed as a type of criterion-referenced test and not as a device to rank the individual members of the group. The rating system is adequate to make valid discriminations within the entire range of proficiency, but not within individual levels. If 50 people are assigned a *2* rating, there is no doubt that the members of that group will differ among themselves with regard to their proficiency. The Oral Interview is not a fine enough instrument to measure individual differences beyond the 11 points on the proficiency scale. As a placement test in a university language program or as a progress test in a language course, the Oral Interview can be very valuable, but it would be unwise to attempt to administer it several times during the same semester.

The matter of test reliability is probably the most critical problem related to the Oral Interview. Even as a high performance automobile needs to be tuned periodically in order to keep it in good condition, so must even the best Oral Interview examiner be checked on a regular basis in order to make certain that he is making judgments that are consistent with the standard criteria which have been established. This means that each examiner must undergo an in-service training period occasionally; otherwise, his judgment in assigning ratings might have a tendency to drift. This "tune up" can be accomplished in 1 of 2 ways. Either all examiners can attend a seminar in which the standards are reviewed, examples of language performance at each level are demonstrated, and participants are given an opportunity to ask questions and exchange experiences, or each tester can listen individually to a set of calibrated tests on tape which have been rated in accordance with the standard criteria.

In conclusion, it should be emphasized that in spite of the problems mentioned above, the Oral Interview can be a valid, reliable, and very useful instrument for testing spoken language proficiency. The examinee is tested in a manner which closely approximates a real-life situation. Only a small random sample of his language is demonstrated, but it is displayed as naturally as is possible under testing conditions. The test cannot be compromised; thus, there is no need to revise it each year. The time normally dedicated to the construction and validation of new forms can be used to train and retrain examiners. Although the test is time-consuming on the part of the examiners, it requires relatively little time on the part of the examinees. The Oral Interview is flexible. The examiners can cater to the level of each examinee, and adjust the topic of conversation according to background, interest, and professional requirements. The quantity of language spoken by an examinee during an Oral Interview is usually quite large, i.e., the number of different lexical and grammatical entities is significant. The Oral Interview thus provides the evaluator with an excellent opportunity to observe the examinee function in the language very much as he would under normal working conditions.

FOOTNOTES

[1] The Oral Interview is also used regularly in the Central Intelligence Agency's language program and by Educational Testing Service under contract for the Peace Corps. The format of the interview differs slightly at each agency.

[2] At the CIA, both examiners are native speaker/instructors. Each has an equal role in administering and rating the interview. In the ETS/Peace Corps program, only one examiner is used.

[3] The test is also known as the "FSI Interview."

[4] See Appendix 1.

[5] See Appendix 2.

[6] A reliability study involving 50 tests each in French, German, and Spanish administered at the FSI and CIA was recently conducted by Claudia Wilds. Results show that the consistency among the raters is very high within each agency, but somewhat lower across agencies. An earlier study involving only FSI examiners also showed high rater reliability.

[7] See for example Perren, 1968, pp. 112 and 115, and Lado in Jones and Spolsky, 1975, p. 25 (comments made during discussion following the presentation of the paper by John Clark).

[8] See Jones, 1976.

[9] See Appendix 5.

[10] At the CIA, each examiner rates independently without consultation. The ratings are then given to a testing assistant who averages the ratings. For information on independent rating, see Goldberg (1973) and Winkler (1967). See Appendixes 3 and 4 for information on the CIA rating system.

APPENDIX 1: Rating Scales for Foreign Language Proficiency
Foreign Service Institute

A. *Original Definitions (1956)*

S-0 No practical knowledge of the spoken language.

S-1 Able to use greetings, ordinary social expressions, numbers, ask simple questions and give simple directions (sufficient for routine travel requirements).

S-2 Able to satisfy both routine social demands and limited office requirements.

S-3 Sufficient control of the structure and adequate vocabulary to handle representation requirements and professional discussions within one or more special fields.

S-4 Able to use the language fluently on all levels pertinent to the Foreign Service.

S-5 Spoken command of the language equivalent to that of English.

B. *Revised and Expanded Definitions (Current)*

S-0 No practical speaking proficiency

S-1 Elementary proficiency
Able to satisfy routine travel needs and minimum courtesy requirements. Can ask and answer questions on topics very familiar to him; within the scope of his very limited language experience can understand simple questions and statements, allowing for slowed speech, repetition, or paraphrase; speaking vocabulary inadequate to express anything but the most elementary needs; errors in pronunciation and grammar are frequent, but can be understood by a native speaker used to dealing with foreigners attempting to speak his language; while topics which are "very familiar" and elementary needs vary considerably from individual to individual, any person at this level should be able to order a simple meal, ask for shelter or lodging, ask and give simple directions, make purchases, and tell time.

S-2 Limited working proficiency
Able to satisfy routine social demands and limited work requirements. Can handle with confidence but not with facility most social situations including introductions and casual conversations about current events, as well as work, family, and autobiographical information; can handle limited work requirements, needing help in handling any complications or difficulties; can get the gist of most conversations on non-technical subjects (i.e., topics which require no specialized knowledge) and has a speaking vocabulary sufficient to express himself simply with some circumlocutions; accent, though often quite faulty, is intelligible; can usually handle

elementary constructions quite accurately but does not have thorough or confident control of the grammar.

S-3 Minimum professional proficiency
Able to speak the language with sufficient structural accuracy and vocabulary to participate effectively in most formal and informal conversations on practical, social, and professional topics. Can discuss particular interests and special fields of competence with reasonable ease; comprehension is quite complete for a normal rate of speech; vocabulary is broad enough that he rarely has to grope for a word; accent may be obviously foreign; control of grammar good; errors never interfere with understanding and rarely disturb the native speaker.

S-4 Full professional proficiency
Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

S-5 Native or Bilingual Proficiency
Speaking proficiency equivalent to that of an educated native speaker. Has complete fluency in the language such that his speech on all levels is fully accepted by educated native speakers in all of its features, including breadth of vocabulary and idiom, colloquialisms, and pertinent cultural references.

APPENDIX 2: Check List of Performance Factors
Foreign Service Institute

1.	ACCENT	foreign	_____:_____	_____:_____	_____:_____	native
2.	GRAMMAR	inaccurate	_____:_____	_____:_____	_____:_____	accurate
3.	VOCABULARY	inadequate	_____:_____	_____:_____	_____:_____	adequate
4.	FLUENCY	uneven	_____:_____	_____:_____	_____:_____	even
5.	COMPREHENSION	incomplete	_____:_____	_____:_____	_____:_____	complete

Instructions for Use of Check List to Determine *S*-Ratings

	1	2	3	4	5	6	(A)
Accent	4	3	2	2	1	0	
Grammar	6	12	18	24	30	36	
Vocabulary	4	8	12	16	20	24	
Fluency	2	4	6	8	10	12	
Comprehension	4	8	12	15	19	23	
						Total	

Procedure: Place in Column (A) the credits to be given for each scale on the Check List. For example, a check mark in position 3 on the "Accent" scale is given a credit of 2. Add the credits to find the total score. The final *S*-Rating is to be equated with the total score by the following table.

Score	Rating	Score	Rating	Score	Rating
16-25	*S-0+*	43-52	*S-2*	73-82	*S-3+*
26-32	*S-1*	53-62	*S-2+*	83-92	*S-4*
33-42	*S-1+*	63-72	*S-3*	93-99	*S-4+*

If the total scores by different examiners yield different *S*-ratings, an average of the total scores should be used to determine the final ratings.

Accent

1. Pronunciation frequently unintelligible.
2. Frequent gross errors and a very heavy accent make understanding difficult, require frequent repetition.
3. "Foreign accent" requires concentrated listening and mispronunciations lead to occasional misunderstanding and apparent errors in grammar or vocabulary.
4. Marked "foreign accent" and occasional mispronunciations which do not interfere with understanding.
5. No conspicuous mispronunciations, but would not be taken for a native speaker.
6. Native pronunciation, with no trace of "foreign accent."

Grammar

1. Grammar almost entirely inaccurate except in stock phrases.
2. Constant errors showing control of very few major patterns and frequently preventing communication.
3. Frequent errors showing some major patterns uncontrolled and causing occasional irritation and misunderstanding.
4. Occasional errors showing imperfect control of some patterns but no weakness that causes misunderstanding.
5. Few errors, with no patterns of failure.
6. No more than two errors during the interview.

Vocabulary

1. Vocabulary inadequate for even the simplest conversation.
2. Vocabulary limited to basic personal and survival areas (time, food, transportation, family, etc.)
3. Choice of words sometimes inaccurate; limitations of vocabulary prevent discussion of some common professional and social topics.
4. Professional vocabulary adequate to discuss special interests; general vocabulary permits discussion of any non-technical subject with some circumlocutions.
5. Professional vocabulary broad and precise; general vocabulary adequate to cope with complex practical problems and varied social situations.
6. Vocabulary apparently as accurate and extensive as that of an educated native speaker.

Fluency

1. Speech is so halting and fragmentary that conversation is virtually impossible.
2. Speech is very slow and uneven except for short or routine sentences.
3. Speech is frequently hesitant and jerky; sentences may be left uncompleted.
4. Speech is occasionally hesitant, with some unevenness caused by rephrasing and groping for words.

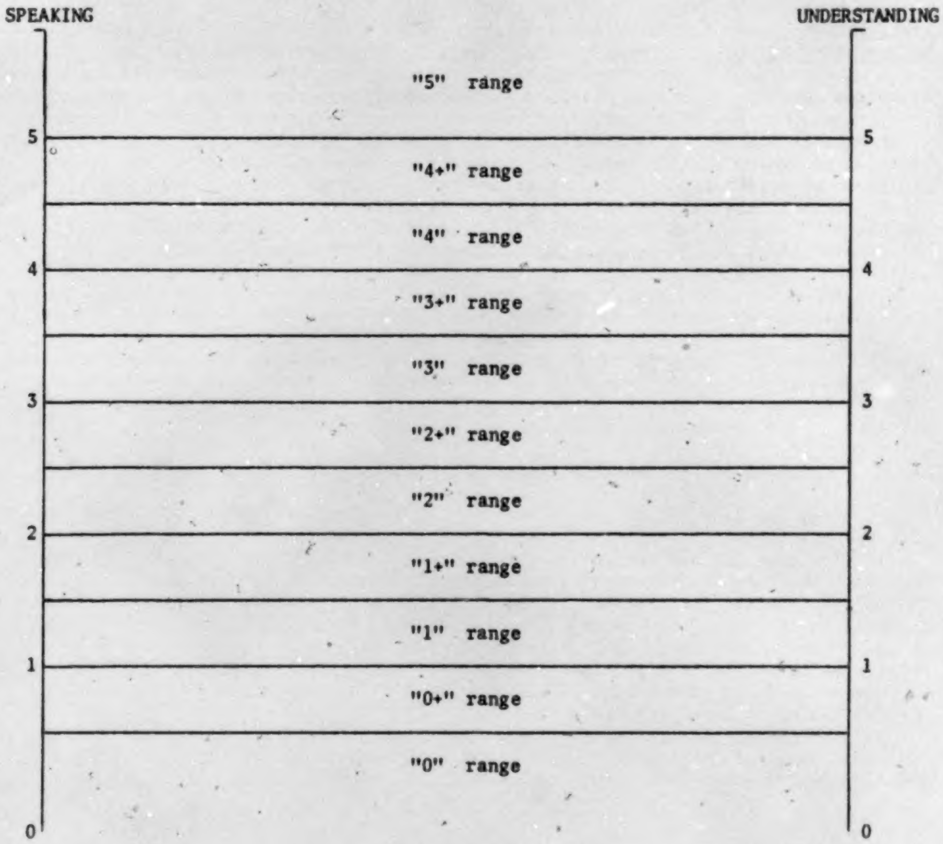
5. Speech is effortless and smooth, but perceptibly non-native in speed and evenness.

6. Speech on all professional and general topics as effortless and smooth as a native speaker's.

Comprehension

1. Understands too little for the simplest type of conversation.
2. Understands only slow, very simple speech on common social and touristic topics; requires constant repetition and rephrasing.
3. Understands careful, somewhat simplified speech directed to him, with considerable repetition and rephrasing.
4. Understands quite well normal educated speech directed to him, but requires occasional repetition or rephrasing.
5. Understands everything in normal educated conversation except for very colloquial or low-frequency items, or exceptionally rapid or slurred speech.
6. Understands everything in both formal and colloquial speech, to be expected of an educated native speaker.

APPENDIX 3: Language Proficiency Ratings (Oral-Aural Skills)
Central Intelligence Agency



APPENDIX 4: Guidelines for Assigning Language Proficiency Ratings
Central Intelligence Agency

0 The examinee has no practical speaking proficiency. He may know a few isolated words and phrases which are of no practical use.

1 The examinee has the minimum proficiency for survival on a day-to-day basis in the target country, i.e. function in simple question and answer situations. He knows enough at this level to satisfy ordinary courtesy requirements. He is able to ask and answer questions relating to situations of simple daily life and routine travel abroad. The examinee is also able to handle requests for services such as getting a hotel room and ordering a meal.

The examinee at this level normally makes errors even in structures which are quite simple and common. His vocabulary is limited to the type of situations mentioned above, and even in these situations he sometimes uses the wrong word. Although his pronunciation may be poor, he makes the minimum contrastive distinctions, including stress, intonation and tone patterns, necessary to make himself understood.

2 The examinee is able to talk in some detail about concrete subjects such as his own background, some current events, his family, travel, experiences, recreational activities, educational background, and familiar places.

The examinee has enough control of the language to be able to join sentences in limited discourse. He has good control of the morphology of the language (in inflected languages), and of the most frequently used syntactical structures. Although his vocabulary is sufficient to talk with confidence about the type of topics described above, his limited vocabulary fairly often reduces him to verbal groping, or to momentary silence. A foreign intonation and rhythm may still be dominant.

3 The examinee is able to converse and express his opinions about topics such as current events, including politics and social problems of a general nature.

The examinee has good control of grammar, though he makes occasional errors in low-frequency structures and in the most complex frequent structures. His vocabulary is broad enough so that he rarely has to grope for words in discussing the topics mentioned above. A foreign pronunciation, though apparent, is not distracting.

4 Although the subject matter that the examinee is able to handle at this level may not differ very much from that of level "3," he is able to use the language in all non-technical situations, and can express his opinions almost as he would in his native language (assuming that he is "5" in his native language). The examinee is able to tailor his speech to his audience. He has nearly perfect grammar and speaks the language with extensive and precise vocabulary. Although the examinee may still have an accent, he very rarely mispronounces the language.

5 The examinee is able to use the language in all conceivable non-technical and semi-technical subjects in a manner equivalent to that of an educated or well-informed native speaker of the language. He may still have a detectable accent, although it interferes in no way with communication.

Assignment of "+" Rating in the Evaluation of Speaking Proficiency

A base level (0, 1, 2, 3, 4) can be modified by a "plus" (+) in any of the following situations:

1. The examinee was considered for a particular base level at some point during the test, and then the examiners decided that he fell short of that level. *Example*: At times during the test Peter "Lee" was "2," but he frequently lapsed into errors ascribed to level "1." (He should then be assigned "1+.")
2. The examinee fulfills the vocabulary requirements of the next higher base level, but falls short of that level in grammar. *Example*: Mary "Lee" is "3" in vocabulary, but her grammar does not exceed "2." (She should then be assigned "2+.")
3. The examinee fulfills the grammar requirements of the next higher base level, but falls short of that level in vocabulary. *Example*: John "Lee" is "3" in grammar, but his vocabulary does not exceed "2." (He should then be assigned "2+.")

APPENDIX 5: Information for Examinees (Excerpts)
Foreign Service Institute

Nature of the Test

A test normally lasts half an hour. The examiners are a scientific linguist and an instructor who is a native speaker of the language involved. The test primarily consists of a conversation with the native speaker on a variety of topics, both social and professional. It may also comprise imaginary situations (such as renting an apartment) and interpreting problems.

Bases for Rating

The examiners consider five main aspects of performance, here listed in the order of their importance: control of grammatical structure, completeness of comprehension, breadth and precision of vocabulary, fluency (or effortlessness of speech), and accuracy of pronunciation and intonation.

Suggestions

1. Answer questions as fully as possible; help us to get the largest sample of your speech that we can. We do not want brief, concise answers; we infinitely prefer loquacity to profundity.
2. On the other hand, try to speak accurately; don't sacrifice grammar for speed.

A Review of the Bilingual Syntax Measure

Ellen J. Rosansky, San Jose State University

With the growing interest in bilingual education, there has been an increasing need for adequate measures of language dominance and proficiency for limited English-speaking children. Silverman et al. (1976) underscore the importance of such evaluation measures:

> One of the most rapidly growing areas within contemporary education is the development of bilingual/bicultural curricula as a response to the needs of limited-English speaking children. In order to properly assess these needs, school districts must determine the language status of these children through some type of equitable evaluation processes. Because bilingual testing for language dominance and/or proficiency is an innovative process for most school districts, it is imperative for them to know the quantity and quality of existing and pilot tests available (p. 5).

Silverman et al. estimate there are approximately 350 Federally funded bilingual education programs in the U.S., but only about three percent of the children who are eligible for some sort of bilingual education program are actually receiving benefits. The lack of adequate language assessment instruments is a factor affecting the implementation of these programs.

There are good reasons why we lack adequate language assessment instruments for limited-English speaking children. Until recently, most of the foreign language proficiency tests were developed for adults, in conjunction with military, government, or university foreign language learning programs. Many of the language proficiency tests in use today, including foreign language tests developed by the Educational Testing Service (ETS) and those used in various English as a second language (ESL) and bilingual programs, have evolved from earlier government language tests. Silverman et al. point out that these tests were designed to "reflect the general language pattern of adults (p. 18)." They question whether such tests could accurately assess child language proficiency, since the phonological, morphological, syntactic, and semantic proficiencies of children may not be the same as in adult language.

While there is no one test which assesses the language proficiency of limited-English speaking children and has the stature and wide acceptance of the TOEFL Test, for example, there are several tests which are intended to measure the oral language proficiency of bilingual children. One such test is the Bilingual Syntax Measure (BSM), developed By Burt, Dulay, and Hernandez-Chavez (1975). The test has been commercially available from Harcourt Brace Jovanovich since 1975. It was designed to measure both the language dominance and the oral proficiency in English or Spanish of limited-English speaking children in grades K-2, in order to aid in the placement of these children in appropriate bilingual education programs.

The Bilingual Syntax Measure was designed both as a measure of syntactic development and as a test aimed at children. The author's goal was to develop a test that is as culture fair as is possible, with items chosen to minimize the cultural and experiential differences among children. Careful attention was given to selecting pictures for the test, and items relating to particular regional or cultural groups have been avoided.

TEST ADMINISTRATION

The test, which takes approximately 10-15 minutes to administer, consists of 25 prepared questions related to seven cartoon-type pictures. It can be administered to individual children by anyone who speaks and writes the language being tested, usually the classroom teacher, an aide, or a school administrator. Administration of the test is fairly simple. As the child is shown each picture, the tester asks some preliminary questions to make sure the child understands the vocabulary items which will be used in subsequent questions, e.g.

> Do you see a fat man? Show him to me.
> And show me the skinny man.
> And the baby birds up in the tree?
> (Point to flowers) And what are those?

The child's responses to these questions are not recorded or scored; they are intended to make sure that vocabulary problems will not contaminate later performance.

After the preliminary vocabulary questions, the child is asked to respond orally to a number of questions about each picture. This approach to eliciting speech is termed "structured conversation" by the test developers. Some examples of the type of questions are:

1. (point to baby birds) What are those?
2. (point to Mama bird and worm) What's the Mama bird going to do with the worm? (this response not recorded)
3. (point to baby birds) Why do they want food?

The authors suggest that this procedure is like chatting with a child about some pleasant pictures. After five questions about the first picture, the examiner is instructed to discontinue the test if the child has not responded to at least three. If the child answered at least three questions, the examiner goes on to the next picture. For each picture, a few preliminary questions are asked to ensure vocabulary comprehension, and then the test questions are asked, with the appropriate picture in full view of the child. The examiner writes the child's responses down in the accompanying test booklet. Next to each line for the written form of the child's responses is either a small square or circle for later scoring, e.g.

6. Point to BOTH HOUSES using whole hand to point
WHAT ARE THESE?________________________________ 6. ☐

8. Point to DOORS OF BOTH HOUSES AT ONCE
AND THESE? ________________________________ 8. ○

There is no tape recording made, although the test manual instructs the examiner to be especially careful about recording errors in a phonologically correct manner and avoiding spelling mistakes (e.g., recording its for it's).

SCORING PROCEDURES

The BSM is designed to yield proficiency levels in either English (BSM-E) or Spanish (BSM-S). The examiner can determine these proficiency levels fairly quickly. If the child has answered fewer than three of the first five questions in any language, the test is discontinued and the child is assigned to Level 1 proficiency. If the child answers at least three questions in any language, but fewer than 13 questions in the language of the test, s/he is assigned to Level 2. If more than 13 questions are answered in the language of the test, the examiner goes on to consider Level 5 proficiency. To do so, s/he scores the eight test items marked by a square in the answer booklet. If the child scores six grammatically correct on these eight designated items, s/he is placed in Level 5 proficiency. If five or fewer of these items are grammatically correct, the scorer then considers an additional 10 items marked by a circle in the test booklet. If seven or more of these items are correct, the child is placed in Level 4; if fewer than seven are correct, the child is placed in Level 3.

The entire test contains 25 questions related to the pictures. Of these, 20 are recorded and 18 scored. The test is thus simple to administer and also simple, in theory, to score. For example, a Level 5 child would have answered at least 13 questions in English (on the BSM-E). The examiner simply turns to the eight "Level 5" questions (marked by a square), and scores those for grammatical correctness. If six of these eight items are correct, the child is placed in Level 5, and the scorer need not look any further.

The BSM is constructed to test syntax; it does not score vocabulary, pronunciation, or functional uses of language. According to the developers, "After careful consideration...syntax was selected as the most suitable basis for the construction of a language development measure intended for general use." The choice of syntax was deemed most suitable because "syntax shows the least variation among speakers of a particular language (*Technical Handbook (TH)*, p. 10)."

The test developers tell us that the "structured conversation" technique makes it "easy to tell what [syntactic] structures children offer in their responses, even when part of a response is absent or misformed (*TH*, p. 14)." In answer to the question, Why does he live here, the answers

(1) He belongs here, or
(2) Because he's fat

are both clearly understandable and grammatically correct. Answers such as

(3) He live here, or
(4) Because he fat

are understandable, and the contexts for an absent 3rd person singular -s (example 3) or copula is (example 4) are clear. Since the test was designed to be as similar to natural discourse as possible, an equally acceptable response would be

(5) Because.

The examiner is instructed to treat responses as acceptable but not scorable in terms of structural errors.

While the scoring method and assignment to proficiency levels are procedurally quite direct, the underlying rationale for the choice of structures tested and assignment to proficiency levels merits detailed examination. The determination of structures and proficiency levels for the BSM is based on earlier scoring procedures adapted by Dulay and Burt (1973; 1974) from child language acquisition studies.

DEVELOPMENT OF SCORING PROCEDURES

While the scoring of the BSM is quite simple, the rationale and development of the current scoring procedure are more complex. Since assignment to proficiency levels and, ultimately, the bulk of the test's subsequent claim of construct validity are intricately intertwined with several aspects of the development of the scoring procedures, it is necessary to examine the earlier procedures carefully. The major areas to be considered here are "structural scoring" and the "syntax acquisition index" (SAI). The interrelation of these two techniques applied to data elicited by an earlier version of the BSM, and analyzed by a statistical method designed to yield hierarchical clusters from dichotomous data, led to the current scoring procedure called the "hierarchic scoring method" (HSM).

Structural Scoring

The procedure called "structural scoring" by Burt et al. is the method developed by Brown et al. (1969) and Brown (1973) in a longitudinal study of three children acquiring their native language. Brown et al. observed that early child language tends to look "telegraphic," i.e. content words (usually uninflected) are supplied. As cognitive growth proceeds, children begin to gradually supply functors, such as prepositions, and the various inflectional morphemes. In seeking a way to characterize the children's linguistic development, Brown et al. observed that certain functors and inflections were often obligatory; e.g., in the utterance Two book, the presence of two creates an obligatory occasion for the plural -s morpheme. Brown (1973) discovered that he could describe the children's growing linguistic abilities by examining the percentage of time each of 14 morphemes was correctly supplied where obligatorily required in each speech sample.

Brown developed a set of criteria for scoring these morphemes. A morpheme was considered "acquired" by the children if it was supplied correctly where required (1) 90 percent of the time, (2) for three consecutive samples, and (3) where there were five or more obligatory occasions for that morpheme. Brown found that as the children were developing over the three years of the study, the ratio of those supplied correctly where required for the 14 morphemes increased in an ordered fashion. The morphemes were rank-ordered according to the order they were "acquired." The order of morphemes for three children was quite similar, and thus Brown reported a sequential acquisition of these 14 morphemes for children acquiring English as their native language.

Burt et al. (see also Dulay and Burt, 1973) used an adaptation of this approach for the data they collected. The major modifications were that (1) they gave "partial credit" if a morpheme was supplied but misformed (e.g., two childs), (2) they retained the 90 percent correct requirement but did not require a minimum of five occasions for each morpheme for which they scored, and (3) they used this approach on cross-sectional data (i.e. one-time data collection) elicited by the BSM, as opposed to the spontaneous, naturalistic speech samples in Brown's study. Since the analysis was based on only one-time samples of speech for each child, they could not include Brown's criterion of "3 consecutive samples." "Structural scoring" for Burt et al. entails scoring the structures elicited by the BSM as: 0/2-no structure supplied where required, 1/2-misformed structure supplied, and 2/2-correct structure supplied. This scoring procedure was used in the early experimental development of the BSM and is still used when the BSM is used as a research instrument.

Syntax Acquisition Index (SAI)

The Syntax Acquisition Index (SAI) entails considering each utterance of the child and computing a ratio. The denominator, called the "Developed Form Value" (DFV), is computed using weighted morpheme values (1,2, and 3 points) for each element of the utterance that should obligatorily have been in the utterance. The numerator, or "Child Response Value" (CRV), is computed using the same weighted morpheme values for tallying what the child actually says. A ratio for each utterance is then computed as in the following example:

```
1        2  1    1  1
Because he are sleeping     (CRV=6)

                                        6/7=.86.

1        2  2    1  1
Because he is sleeping      (DFV=7)
```

To calculate the child's SAI a new ratio is computed where the numerator is the sum of all the CRV's (six in the above example), and the denominator is the sum of all the DFV's (seven in the above example). The ratio is then multiplied by 100. Thus, the SAI=ΣCRV/ΣDFV x 100.

SAI and Proficiency Level Assignments

The SAI is the source of the later proficiency level assignments for the BSM, although this evolution is also somewhat complex. The SAI relates to proficiency Levels 3,4, and 5 in the following manner. An SAI of 95-100 was labeled Level 5, 85-94 was considered to be level 4, and 45-84, Level 3. Such a distribution of SAI ranges into proficiency levels results from the following analysis.

In establishing the cut-off points for proficiency levels, Burt et al., observed that 75 percent of the children in Level 3 had scores of 90 percent or better on two structures, 75 percent of those assigned to Level 4 scored 90 percent or better on nine structures, including the two Level 3 structures, and 75 percent of those children assigned to Level 5 scored 90 percent or better on five additional structures on the BSM-E.[1] Similar decisions were made for the BSM-S.

Burt et al. suggest that this sort of proficiency level assignment is validated by other research in second language acquisition. According to them, "the grouping of structures acquired at each proficiency level is similar to that reported in previous basic research on English acquisition order (*TH*, p. 19)." Here, they refer to their own research using the BSM as a speech elicitation instrument, where they claim that there is an ordered sequence in the morpheme acquisition of second language learners. They also cite several other studies in their support here, notably Bailey et al. (1974) and Krashen et al. (1975), where the BSM was used as an elicitation instrument with adult ESL students, and the group percentage correct for eight of the same structures was rank-ordered. The rank order was significantly correlated with the order Dulay and Burt obtained using the BSM on children.

Hierarchical Scoring Method (HSM)

The current scoring technique for assignment to proficiency Levels 3, 4, and 5 was arrived at after one additional change. Initially, Dulay and Burt (1973; 1974) reported an ordering or "acquisition sequence" of morphemes in the research literature based on data elicited by the BSM. However, the ordering was "not sufficiently precise (*TH*, p. 21). The lack of precision was attributed by Burt et al. to the lack of an adequate technique for specifying "more meaningful relationships among structures (*TH*, p. 21)." Other researchers in the field (e.g., Larsen, 1975a and b; Rosansky, 1976a and b, 1977) have questioned whether there is any reason, linguistic or cognitive, to expect these structures to be ordered with respect to each other, in terms of sequential acquisition among second language learners.

Ordering Theoretic Method (OTM)

Burt et al. employed a technique developed by Bart and Krus (1973) in re-analysis of the earlier "structural scoring" data. The OTM depends on dichotomized data. In this case, a score of 1 was assigned to a structure if it attained 90 percent or better and 0 if the score on that structure was 89 percent or less. This method produces a hierarchical arrangement of the data. Usually, implicit in the use of such a technique is an *a priori* hypothesized hierarchy. The analysis considers paired ordered data, indicating the percentages of disconfirming ordered pairs. If there are few cases disconfirming the ordering, the hypothesized order (e.g., article precedes plural) is then considered to be established. Burt et al. explain what they mean by few cases by saying that, "one structure was said to precede another if five percent or fewer disconfirming cases were found, although occasionally six percent was accepted...(*TH* p. 23)."

After the ordering among pairs of structures is determined, a hierarchy which incorporates all such pairs is arranged. Here 18 structures are listed for analysis rather than the earlier mentioned 14 structures. Such structures as the 3rd person copula and auxiliary divided into two categories each, singular and plural, and the components of the earlier conditional perfect were divided into the conditional auxiliary, perfect auxiliary, and past participle. Each is analyzed as a separate structure in the ordered pairs and hierarchical analysis.

The end product of this analysis is a collection of morphemes clustered together and unordered with respect to each other, but ordered with respect to certain other clusters of morphemes. Prior ordering among clusters is interpreted as acquired earlier. According to Burt et al., "the structures near the top of the hierarchy are those...demonstrated to be associated with Level 5. The structures associated with Level 3 precede all the others, while the Level 4 structures fall in between (*TH* p. 23)." They take the results of this hierarchical analysis as additional support for an ordering where presence of higher level structures implies previous acquisition of lower level structures and "the absence of some lower-level structures implies the absence of the upper-level structures (*TH*, p. 21)."

An examination of the graphed data and tables presented in the *Technical Handbook* leaves one wondering, however, about the relative lack of ordering of so many structures (four or five morphemes can be in the same cluster) when the earlier structural scoring analysis yielded a rank ordering among these very same morphemes. Furthermore, while the hierarchical analysis offers attractive possibilities, no explanation or rationale has ever been offered either *a priori* or *a posteriori* for the actual clustered ordering that emerges from this analysis. And thus we are no closer to specifying "more meaningful relationships among the structures (*TH*, p. 21)."

Ordering Theoretic Method and the Hierarchic Scoring Method (HSM)

The obtained implicational series just described forms the basis of both the current scoring procedures and the 18 BSM test items. The higher level structures from the hierarchic analysis comprise the items elicited by the Level 5 questions, and the lower level structures comprise the Level 3-4 items. In other words, the eight test items designated as Level 5 items on the BSM were designed to elicit the four structures identified by the Ordering Theoretic Method as higher level structures (i.e. possessive 's, long plural -es, past irregular, and conditional perfect), and the 10 remaining questions were designed to elicit lower level structures. Thus, based on the Ordering Theoretic Hierarchy, correct answers to Level 5 items implies mastery of the other lower level structures.

SAI and HSM: A Comparison

From the perspective of ease of scoring, the current scoring technique (HSM) is procedurally simplified; it does not require making decisions about the obligatory occasion of each structure, but only judgments about the grammatical correctness of the entire response. The final form of the test elicits data on nine of the original structures.[2] Some of the items elicit more than one morpheme in the typical response. In a sense some information is lost with the HSM when the same utterance contains occasions for more than one morpheme, since if any part of the utterance is not grammatical, the response is scored as ungrammatical. For example, an item such as:

> (point to the eyes of both green fish)
> 14. Why do you think their eyes are closed?

was designed to elicit -ing and the auxiliary be, or the copula be. Appropriate responses might be:

(1) Because they are sleeping.
(2) Because they are tired.

It is also possible to respond:

(3) Because their eyes hurt.

Example 3 includes none of the hoped for morphemes, although it does include the plural -s morpheme. This response would be scored as correct using the HSM, but since only one or two other items in the test are intended to elicit the desired structure, the child has fewer opportunities to demonstrate his mastery of it.

Consider another response to item 14 above:

(4) Because he are sleepihg.

With HSM this response is scored as ungrammatical and receives a zero. Responses 1-3 receive a score of 1. In the SAI framework, response 4 would be scored as .86.[3] So in this way, the HSM gives the child less credit for his syntactic ability than the SAI. The two methods are clearly different, and while the SAI has the advantage of greater sensitivity to the child's syntactic ability, the HSM has the advantage of simplicity.

Burt et al. acknowledge that using the HSM results in lower proficiency level assignments than the SAI. A comparison of the two tests, based on scores from 1275 children, reveals that on the BSM-E, 11.3 percent of the children were rated at a higher level using the HSM instead of the SAI, while 20.8 percent were placed in a lower proficiency level when scored on the HSM rather than the SAI. For the BSM-S, 5.8 percent were classified at a higher level, and 20.7 percent were placed in a lower level with the HSM. Overall, there is a 68 percent agreement between level assignments obtained using the SAI and HSM on the BSM-E, and a 74 percent agreement on the BSM-S. Put another way, 32 percent of the time both scoring methods do not yield the same proficiency level assignments for the same child on the BSM-E, and the proficiency level assignments are not the same for the BSM-S 26 percent of the time.

An examination of the two methods, according to Burt et al., "suggests a strong relationship between the SAI and HSM (*TH*, p. 29)." This is to be expected, since the HSM was derived from the earlier SAI. In another comparison of the two scoring methods, they list each possible score for Level 5 and for Level 3-4 items based on the HSM, and then compare these to the mean SAI scores for the children obtaining these HSM scores. They report correlations for the BSM-E of .78 (Level 5) and .75 (Level 3-4). They do not indicate what kind of correlations these are, or if they are significant.

Practicality of Application

With the exception of the reports of various second language researchers who have used the BSM to elicit speech samples, no first hand reactions from classroom teachers administering the BSM have been reported. Thus, it is difficult to assess how simple and efficient the average classroom teacher actually finds the test to be for purposes of administration and scorability.

In a technical report from the Culver City Project, a bilingual immersion program, Lebach (1975) reports on the administration of the BSM to 12 Anglos and six native Spanish speakers. An earlier version of both the BSM-E and BSM-S were administered to both groups of children. This edition contained 33 items rather than the current 20, and it utilized the SAI scoring.

The report indicates that 100 percent of the Anglos and 90 percent of the native Spanish speakers scored at Level 4 or 5 on the BSM-E, and all of the children were placed in Levels 4 or 5 on the BSM-S. The report notes that there was, at the time, no information regarding the establishment of the different levels. Happily, the current *BSM Technical Handbook* provides a detailed explanation of the establishment of the SAI and HSM proficiency levels.

Lebach reports some of the problems encountered in administering the test. One problem was the difficulty in recording the child's response accurately, even though there were two people writing down the responses.[4] The current test instructions require that only the test administrator record the child's response. Lebach observes that while the test takes only 15 minutes to administer prior to scoring, this still adds up to a great deal of time, since each child must be tested individually. She questions whether teachers would be willing to spend the time recording the responses.

A number of problems were also encountered during the scoring (using the SAI). Lebach found that the test took a long time to score. She also found many responses unscorable and was unclear about how to deal with these. Some of these problems have been corrected by the current scoring techniques, which are fairly simple. The *Manual* accompanying the published version of the BSM also attempts to anticipate many of the scoring problems and gives specific examples and appropriate procedures.

RELIABILITY

Burt et al. report two measures of reliability--test-retest reliability and inter-scorer reliability. Test-retest reliability is reported for the BSM not on the actual scores the children received, but on the stability of the proficiency level classifications in the retest. The reliability sample consisted of 154 Spanish speaking children in grades K-2 from 11 schools in two regions of California. The BSM was administered twice in English, followed by twice in Spanish or the reverse, with the order assigned randomly. The two tests in each language were administered approximately two weeks apart and by different examiners at each administration. The tabled data reveal the BSM-E reliability sample consisted of 147 children, and the BSM-S sample consisted of 143 subjects.

The test-retest reliability figures are reported in terms of proficiency levels obtained on the BSM. Burt et al. suggest that the data do not meet the assumptions necessary for calculating parametric (Pearson product moment) correlation coefficients, noting that when assignment to all five levels is considered, the data are discontinuous, since Level 1 and 2 are qualitative rather than quantitative judgments about language proficiency. The subjects were also unevenly distributed across the five proficiency levels, with a disproportionate number classified as Level 4 or 5.

The BSM test-retest showed a 66.1 percent agreement between proficiency level assignments in two administrations of the BSM-E and 76.2 percent agreement on the BSM-S. Chance agreements would be 38.8 percent and 64.5 percent respectively. Rather than the more traditional reliability measures, the Kappa coefficient of agreement is reported as a measure of reliability. Both weighted and unweighted Kappas are given for the test, with K=.45 for the BSM-E and K=.33 for the BSM-S. Kappa can be interpreted as the proportion of agreement after chance agreement has been excluded. The Kappa coefficient is also reported for the language dominance classification, and the authors conclude that this is about as reliable as either language separately. As reported in the *Technical Handbook*, a calculation of Kappa for Levels 3-5 only, on the BSM-E, yields a K=.26 compared with K=.45 for all five levels of the BSM-E.

Burt et al. claim "Kappa coefficients [reported by Cohen, 1960; 1968]....Are intended for use in assessing the reliability of classifications into nominal or *ordinal* [italics added] categories (*TH*, p. 40)." An examination of Cohen's explanation of Kappa reveals three conditions which, he proposes, "may be taken as assumptions of the coefficient of agreement (Cohen, 1960, p. 38)." Cohen describes the measure as one dealing with *nominal* data, e.g., clinical psychologists categorizing patients into such categories as schizophrenic, neurotic, and brain-damaged on the basis of different units or psychological tests. Accordingly, then, the conditions or assumptions for the use of Kappa are:

(1) the units are independent,
(2) the categories of the nominal scale are independent, mutually exclusive, and exhaustive, and
(3) the judges operate independently,

where the reliability concern is with degree of agreement between two judges, and "in the typical situation, there is no criterion for the 'correctness' of judgments (Cohen, 1960, p. 38)."

It is not clear why the authors chose to report the reliability of essentially ordinal data in terms of a measure designed for use on nominal data. This is particularly unclear when they have recourse to the actual syntax measures in addition to the proficiency level assignments. Moreover, it is not clear that ordered proficiency levels ought to be treated as nominal data meeting the assumptions specified by Cohen for the Kappa coefficient. The categories (proficiency levels) are not clearly independent, but instead derive from continuous data for which cut-off points have been established to create categories. Nor are the judgments clearly independent; they depend on a specific subset of linguistic structures, not all equally elicited by the test, for which there is a fairly clear criterion for the "correctness" judgment.

Examination of the data presented in Tables 33-36 of the *Technical Handbook* (pp. 40-42) indicates that for the BSM-E the proportion of agreement in the assignment of proficiency levels is 66.1 percent, including chance agreement. One way to interpret this is that 34 percent of the pupils'tests (50/147) were not assigned to the same proficiency level on both administrations of the test, and 24 percent (34/143) of the students were not assigned to the same proficiency level on the BSM-S. This is particularly disconcerting when we examine the actual distribution of proficiency level assignments. The bulk of students on both the BSM-E and BSM-S were classified as either Level 4 or Level 5 proficiency. Of the 50 students not placed in the same proficiency level on both administrations of the BSM-E, 36 (72 percent) were Level 4/5 combinations (i.e. they were classified as Level 4 on one administration and as Level 5 on the other). On the BSM-S, 29 of the 34 students (85 percent) who were not classified at the same level on both administrations were also Level 4/5 combinations. Since the test is designed to discriminate among children on the basis of their syntactic proficiency so as to place them in appropriate ESL/bilingual classrooms, these reliability problems are cause for concern, particularly if placement decisions rest solely on BSM proficiency level assignments without corroborating evidence from other language assessment measures.

While the test score data from the BSM may not meet the conditions for the computation of a conventional Pearson product moment correlation coefficient, several alternatives are possible.

(1) A standard test-retest reliability coefficient could have been computed for Levels 3-5, since these levels are all determined in the same fashion and the distribution of scores across these three levels more closely approximates a normal distribution, or

(2) If normalcy is still in question, Spearman rank-order correlation coefficients could have been calculated between the relative scores obtained by the reliability sample on the two administrations of the test.

Since the assignment to proficiency levels is directly derived from the students' scores, one could be fairly certain that, if on test-retest the students' relative scores were similar, then their subsequent assignment to proficiency levels would be similarly comparable.

Interscorer Reliability

The second type of reliability measure reported by Burt et al. is interscorer reliability. The authors acknowledge that "there is an element of subjectivity in deciding whether a specific response is or is not grammatical (*TH*, p. 44)." Thus, they felt it was appropriate to assess the reliability of the scoring procedure.

The sample for the reliability study was based on 260 children of mainly Mexican background in grades K-2, taken from two of the Field Test Centers. Since this reliability measure is related specifically to the comparability of two scorers' judgments, those children assigned to Levels 1 and 2 were excluded from the study. The interscorer reliability study used HSM for scoring. The reliability of the two scorers was tested by copying the children's responses for the 18 scorable items in new test booklets. The two persons tested for interscorer reliability were thus not involved in the test administration or the initial scoring process. Their task was to judge already recorded responses, and was thus different from what the teacher actually does when administering the test. The teacher or examiner must, in addition, make judgments during the test itself in order to write down what the child says as accurately as possible.

As with the test-retest reliability, the degree of agreement is reported in terms of percent agreement: 84 percent for the BSM-E and 80 percent for the BSM-S, and then in terms of Kappa. The unweighted Kappa's obtained were K=.73 (BSM-E) and K=.59 (BSM-S). Once again, it is not clear why the interscorer reliability is reported in terms of a measure designed for nominal data.

Summary of Reliability

The two sorts of reliability reported are test-retest--K=.45 (BSM-E) and K=.33 (BSM-S)--and interscorer reliability, also reported in terms of Kappa--K=.73 (BSM-E) and K=.59(BSM-S).

What should be remembered here is that test-retest reliability refers to the assignment to proficiency levels and not to scores themselves. The interscorer reliability refers to the amount of agreement between scorers in the dichotomous rating of 18 responses as grammatically correct or incorrect. Neither scorer was at the test scene.

Even if the authors' methods are accepted, it can be seen that the reliability data reported are low. The authors acknowledge this general lack of high reliability and go on to explain that, "while children are acquiring particular syntactic structures, they will use them correctly on some occasions and incorrectly on others...(*TH*, p. 45)." They suggest that this variability will be reflected in the children's responses to the BSM items and that, "A very high level of reliability for this test would suggest that it was not sensitive to... changes in a child's speech and, hence, not adequately reflective of the acquisition process (*TH*, p. 45)."

Another explanation for the relatively low reliability reported is also possible. There are certain factors which are generally known to affect reliability. In addition, some sources of error variance are more likely to influence reliability in certain circumstances than are others.

(1) A well-known factor affecting reliability is the length of the test. Generally, the longer the test, the better its reliability. The BSM, with only 18 items, is not a long test. If it were longer, reliability would probably be improved.

(2) The variability of the sample can also affect test reliability. All else being equal, higher reliability will usually be obtained for a group which varies more in the ability being tested. The distribution of scores for the field test samples indicates that most of the students were at Level 4 or 5 proficiency, and were thus quite homogeneous. Few of the subjects tested, in fact, were beginning learners of English, 79 percent of those tested were in the U.S. for more than three years, and 62 percent of the field test sample had actually been in the U.S. since birth. Only 7 percent of the children had been in the U.S. for one year or less. Higher reliability might have been obtained if the sample were more heterogeneous in terms of the children's knowledge of English.

(3) Lyman (1978) discusses several sources of error variance. One of the sources he identifies, "examiner-scorer influence," can be, and often is "negligible on standardized paper and pencil tasks (p. 34)." While he combines examiner-scorer as a single category, it is useful here to separate these two sources of error variance. Interscorer reliabilities, as reported above, are not particularly low. But this type of test is more likely to be subject to *examiner* influence. As Lyman observes, "the examiner is likely to be a major source of error variance when the test is administered individually (p. 35)." There can be substantial differences in the test results obtained by different examiners. In addition, errors in recording and copying test responses are also potential sources of error variance.

Thus, low reliabilities may be a reflection of second language in progress, as Burt et al. suggest, but there are a number of other variables that might also be influencing the reliabilities of the BSM.

VALIDITY

The sections which follow will consider various issues related to test validity. These sections also bear directly on research since, as Davies (1977b) observes, "The first aim of language testing is that of research....Whether mother tongue or second language situations are being considered, whether research is being conducted in the laboratory or in natural settings, tests will at some point be necessary to check on the learning that has taken place so far (pp. 41-42)." He goes on to say that "...tests under construction should not be ransacked for experimental results before they have reached the validation stage. Anything else is to make available results that are certainly non-valid and possibly unreliable also (1977a, p. 7)." Thus, the discussion of validity which follows may be taken to apply to the use of the BSM as a research instrument, as well as to its use as a language proficiency test.

Burt et al. discuss validity in terms of construct validity. They suggest that measures of predictive or concurrent validity are not appropriate, claiming that there are no other suitable measures of syntax which would allow comparative studies of the BSM and other instruments.

Construct validity, the authors rightly point out, cannot be established by a few research studies, but depends on the composite evidence available which supports the claim that the test measures the construct it was intended to measure. The BSM was designed to assess the acquisition of ESL and Spanish L_1 (a first language) maintenance, as well as language dominance, for kindergarten through primary grade children as reflected by syntactic proficiency. A review of validity must, therefore, consider whether the BSM measures what it claims to measure, i.e. a language development construct. Furthermore, since construct validity depends upon all available evidence, a discussion of other types of validity is also included.

Construct Validity

The BSM was designed to assess a psycholinguistic construct, that is, language development as indicated by syntactic proficiency. The authors claim that previous research in child and second language acquisition led them to postulate a hierarchical acquisition of syntactic structures in both English and Spanish. Subsequent data collection during the research program (using the BSM) gave further support to the existence of such structural hierarchies and led to the establishment of proficiency levels.

The authors' intention was to develop a scoring method "that would indicate the status of children with respect to specific points along a developmental psycholinguistic continuum (*TH*, p. 14)." Unfortunately, contemporary second language acquisition researchers do not agree about the nature of such a continuum and even about the existence of such invariance at the level of morphology among second language learners. Larsen (1975b) concludes that, "there does not seem to be an invariant ordering of morphemes...(p. 68)." Rosansky (1976a) observes that "while there are surely commonalities among language learners, the analysis of morpheme performance may not be the most productive way to search for...[them] (p. i-ii)." Rosansky goes on to report that morpheme performance, when observed longitudinally in the spontaneous speech of second language learners from one month to the next, tends to fluctuate, and that morphemes as a group do not maintain a stable ordering relative to each other. Her longitudinal data suggests a great deal of variability in individual morpheme accuracy and that possibly by grouping the individual data, this variability is obscured. Finally, particularly with regard to the underlying claim for construct validity based on a "developmental psycholinguistic" construct of language acquisition, Rosansky's work questions whether we can reasonably assume (as does the BSM) that the cross-sectional rank ordering of morpheme accuracy indeed represents acquisition.

The rationale for the BSM proficiency levels is based on the authors' claim that there is "a common order of acquisition of certain English grammatical morphemes by children of diverse language backgrounds and learning environments acquiring English in the United States (*TH*, p. 12)." Burt et al. note that three groups of Spanish speaking children from diverse backgrounds learning English in the United States all had the same "acquisition sequence" for eight grammatical morphemes (Dulay and Burt, 1974). There is a problem in using this sort of evidence of a "natural sequence" in support of a sequential acquisition construct, since the data were collected using an experimental version of the BSM itself. Because the same order was obtained during the field development analysis of the test results, the authors cite this as empirical evidence which demonstrates the construct validity of the test. While they don't actually claim concurrent validity, in choosing and demonstrating construct validity for assignment to proficiency levels, they are in effect doing so on the basis of apparent concurrent validity with data elicited from their own test.

Davies (1977a) addresses just this issue. He notes that "most so-called tests which are not yet standardized are really trial tests, based on various unsubstantiated ideas and guesses (p. 5)." He observes that, "One of the abuses of tests under construction--the ideas, guesses, trials...referred to--is that they are often used as if they were already tests in use and employed in order to validate or assess experiments (p. 5)." Construct validity is evaluated by investigating what psychological qualities a test measures--that is, by determining the degree to which constructs or explanatory concepts account for performance on the test. In the case of the BSM, the construct under consideration is a language development construct which rests on the notion of a natural sequence of morpheme acquisition--or at least on a statistically induced three-tier hierarchy among 8-12 morphemes. It should be recalled that earlier research based on the BSM and scored by the SAI (percentage correct) method produced performance data on specific morphemes for groups of individuals. These data were subsequently re-coded as correct or incorrect and submitted to the ordering theoretic method.

Thus, the construct (i.e. the postulated hierarchic or developmental sequence) appears to account for performance on the test. It should be emphasized, however, that the construct which is being used to account for performance on the test, or to assert construct validity for the BSM, is postulated based on cited examples of earlier research in the field (Dulay and Burt, 1973 and 1974; Bailey et al., 1974; Krashen et al., 1975) using experimental versions of the test itself.

It would appear that conclusions based on data from the test are thus being using to validate later performance on the same test. That is, the effect of the argument for construct validity, particularly in the absence of any other supporting evidence of validity (e.g., content, concurrent, or predictive validity), is circular.

The construct validity of a test is essentially the degree to which a test can be said to measure some theoretical construct or trait. Anastasi (1976) has observed that, "construction validation requires the gradual accumulation of information from a variety of sources [and] any data throwing light on the nature of the trait...and the conditions affecting its development and manifestations are grist for this validity mill (p. 151)." Construct validity is a rather complex sort of validity. According to Thorndike and Hagen (1977), a construct is a term from psychology which "refers to something that is not observable, but is literally 'constructed' by the investigator to summarize or account for the regularities or relationships that he observes in behavior (p. 70)." In claiming construct validity one is essentially theorizing about a trait, and such a theory leads to certain predictions which must be tested. One prediction Thorndike and Hagen point to is that, "the nature of a trait, and therefore of valid measures of it, will indicate that it should be related to certain other measures (p. 71)." The kinds of measures with which it should be correlated may include other, already established measures of the construct under consideration, or ones with which the construct should logically be related. With regard to the BSM, then, we would want to have evidence from correlations of the BSM with other already validated tests of oral syntactic proficiency or with measures of other domains of language acquisition with which syntactic proficiency ought to be moderately related. As we have already observed, such evidence has not been offered.

Another way to test a theory about a trait or construct relates to predictions about group differences. As Thorndike and Hagen point out, "For any given trait, our general knowledge of our society and the groups within it will suggest an array of group differences that seem to make sense. Applying a test to these groups, the investigator finds out how consistently his predictions are borne out (p. 72)." An example of this reasoning applied to the construct of language acquisition upon which the BSM is based would be to speculate about the differences between the acquisition sequences of children and adults, or among learners from diverse native language backgrounds, or between the order of morphemes obtained using the BSM to elicit speech data from L_1 as opposed to L_2 children. We can consider these as three separate predictions. The first speaks to the prediction that children and adults, owing to differences in their cognitive make-ups, learn a second language in a different fashion. That children and adults have different cognitive abilities has been demonstrated in numerous studies within the psychological literature. Several studies within the psycholinguistic literature have suggested that children and adults learn langauge differently as well. A test of the validity of the BSM using this prediction testing approach would be to observe whether the data obtained using the BSM to elicit data from children and adults provides similar or different accuracy orders of the morphemes. The studies conducted by Bailey et al. (1974) and Krashen et al. (1975), cited by Burt et al. in a different context, provides us with some insight into this prediction. Bailey et al. report high correlations ($rho=.88-.94$; $p<.01$) between the order of morpheme accuracy they obtained using the BSM to elicit speech from adult ESL students and the order reported by Dulay and Burt (1973) for young L_2 learners. Rather than question the validity of the language acquisition construct as they have posited it, or of the test as they have developed it, Burt et al. cite the Bailey et al. study as basic research evidence which supports the notion of an invariant sequence of morpheme acquisition.

A second prediction speaks to the differences in language acquisition among learners of diverse language backgrounds. Many researchers and teachers have observed that whatever the overall process of language acquisition may be, at least some learner errors are attributable to native language background. To be sure, while a strong contrastive analysis view does not seem tenable, an error analysis view of language acquisition does provide for the possibility of native language interference. Nevertheless, Dulay and Burt (1974a), in a study based on data elicited by an earlier version of the BSM, claim that interference errors are almost non-existent. Again, rather than question the validity of their instrument, they use these results to support their invariant morpheme sequence notion. In contrast, Larsen (1975) compared the order of morphemes obtained from four of her own tasks and the BSM on adult ESL students from a variety of L_1 backgrounds. She found that, "...Spearman rho's between language groups were often not significant with the exception of the Bilingual Syntax Measure...(p. 9)."

Finally, a third issue speaks to the predicted differences between L_1 children acquiring English as their native Language and L_2 children acquiring English as a second language. When Dulay and Burt (1973; also Dulay, 1974) obtained their initial morpheme accuracy order, they compared their order to the order obtained by Brown and his colleagues (1969) and by Brown (1973) for L_1 children based on spontaneous speech data. The orders did not correlate significantly. Dulay and Burt argued that this was to be expected, since the cognitive abilities of children in the process of acquiring their native language are different from the cognitive abilities of older children acquiring a second language. This is a convincing argument and ought to obtain in a direct comparison between BSM elicited orders of morphemes from L_1 and L_2 children. That is, we would expect that the order of morphemes obtained from L_1 children based on data elicited by the BSM would not correlate significantly with the order obtained by Dulay and Burt, although it should be expected to correlate with the order already reported in L_1 research. Porter (1975) administered the BSM to L_1 children and compared the morpheme accuracy order he obtained with the order reported by de Villiers and de Villiers (1973) for L_1 children (an order which correlated with Brown's order) and with the order reported by Dulay and Burt (1974b) for L_2 children. Porter found that the order he obtained correlated with the Dulay and Burt (1974b) order (rho=.80; $p<.05$), but did not attain statistical significance with the L_1 order reported by the de Villiers. He suggests that the order obtained using the BSM may possibly be an artifact of the BSM testing situation and that the BSM "be subject to more rigorous validation with...the...spontaneous speech production of...second language learners (p. i.)."

In summarizing the above predictions and their impact on the claim Burt et al. make for construct validation, several predictions can be formulated in connection with the proposed construct of language acquisition. As discussed, these predictions are not well borne out, or, more precisely, often it is data elicited by the instrument under consideration which do not seem to support certain of the predictions. In some cases the authors of the instrument have, in fact, used resulting data to differ with certain predictions already previously supported in the literature and to argue their own theoretical views regarding the existence of an invariant morpheme accuracy order.

Thorndike and Hagen explain that, "For any test that presumes to measure a trait or quality, we can formulate a network of theory, leading to definite predictions. These predictions can be tested. Insofar as they are borne out, the validity of the test as a measure of the trait or construct is supported. Insofar as the predictions fail to be verified, we are led to doubt the validity of the test, or our theorizing, or both (p. 73)."

Criterion Related Validity

According to the APA *Standards for Development and Use of Educational and Psychological Tests*, criterion related validity is obtained by comparing test scores (or predictions made from them) with an external variable which is considered to provide a direct measure of the characteristic or behavior in question. Petersen and Cartier (1975) give an example from the Defense Language Proficiency Test (DLPT) as a predictor of real-life language performance. The DLPT yields a score which is the examinee's relative standing on a test which samples representative examples of "language proficiency." They note that even if a test can logically be assumed to sample the relevant content domain, converting scores to skill levels "is not warranted on the basis of content validity alone; to do that, we would have to actually establish the criterion-related validity of the...[test] by correlating [it] with external criteria (p. 107)." They go on to say that, "To perform such a conversion, relevant real-life performance criteria must be collected in quantitative form and the...scores correlated with them to provide predictive validity coefficients and corresponding errors of estimate. The next best thing would be to correlate them with interview ratings...(p. 107). Both predictive validity and concurrent validity can be seen as criterion related types of validity. As already noted, Burt et al. suggest that measures of predictive or concurrent validity are not appropriate for the BSM and therefore provide no evidence of either predictive or concurrent validity.

Predictive Validity

Predictive validity can be seen as information about the practical utility of the BSM proficiency level assignments. Burt et al. observed that during the test's development, most of the children scored at Level 4 or 5 on both the English and Spanish versions. With so little discriminability, it would be desirable to know whether assignment to Level 4 or 5 (or even Level 3) makes any difference in the placement of the child in the appropriate bilingual education program or ESL class. While the Level 1 or Level 2 child has a clear need for bilingual education or second language instruction, what do the other three proficiency level rated children have in common? The question, then, is whether the other three proficiency levels, based directly on the specific syntactic features in the test itself, are actually useful designations. Does it aid the teacher or program administrator to know whether a child attains Level 3, 4, or 5? Are these level assignments useful for class assignments? Do the levels predict the child's further progress in the acquisition of other syntactic features of the language being tested? Do the levels relate, in any systematic way, to the child's proficiency in the vocabulary, phonology, or overall functional use of that language? Do the specific levels predict or relate in some way to general academic progress?

It would be of interest to gather data from school systems which have used the BSM, to establish whether initial assignment to a particular level relates to other aspects of the child's developing communicative competence. In other words, if 10 children are assigned to Level 5, and 10 children are assigned to Level 4 proficiency according to the BSM, are there observable differences between these two groups, either from the subjective impressions of their teachers (and apart from any Pygmalion effect) or in terms of subsequent communicative development? Until such time as this kind of data is available, one must exercise caution in evaluating children on the basis of such proficiency level assignments.

Concurrent Validity

Clark (1975) outlines some theoretical considerations in oral proficiency testing. He observes that, "A direct test of oral proficiency, in the face-to-face communication sense, would involve a test setting in which the examinee and one or more human interlocutors do, in fact, engage in communicative dialogue. A major requirement of direct proficiency tests is that they must provide a very close facsimile or work sample of the real-life language situations in question, with respect to both setting and operation of the tests and the linguistic areas and content which they embody (p. 11)."

Clark distinguishes between the above direct proficiency tests and what he terms indirect proficiency tests which "do not require the establishment of a highly face-valid and representative testing situation. In some cases...an indirect test may involve certain quasi-realistic activities on the students' part. For example...a test which is defined...as indirect may require the student to describe printed pictures aloud..." He points out that, "...since such testing procedures are not truly reflective of a real-life dialogue situation, they are considered indirect rather than direct measures of oral proficiency (p. 11)."

The BSM is generally administered to children by classroom teachers or a school administrator. Although the authors make a claim for the naturalness of the language elicited by the BSM questions, there is reason to suppose that the structured conversation technique is indeed highly structured. Jones (1975) has observed that "most structured oral language tests use a text, pictures or a recording as the stimulus. The response of the examinee is limited and often unnatural. There is little possibility for variation...(p. 4)." The BSM seems to fit the description offered by Jones. It is a structured oral language test. It is also an indirect proficiency test in Clark's terms.

Indirect measures can be useful in sampling aspects of real life events, however. As Clark points out, the validity of "indirect procedures as measures of real-life proficiency [can only be] established through statistical...(correlational) means (p. 11)." An indirect test can be assumed to be a useful measure if it correlates highly and consistently with more direct tests of proficiency. This is what is known as congruent or concurrent validity.

Davies (1977a) discusses the various types of tests and the kinds of validity traditionally associated with each. He notes that usually "achievement and diagnostic tests make use first of content validity, proficiency tests establish predictive or concurrent validity, and aptitude tests claim construct validity. If possible, a second validity type will also be established (e.g. content validity for proficiency tests...(p. 5))." Thus, the BSM, as a proficiency test, would do well to establish some kind of concurrent or predictive validity. In the absence of available predictive validity data, some estimate of concurrent validity would seem to be essential and highly appropriate.

There are several ways to establish concurrent validity. One way is to compare the results obtained using the BSM with results obtained on the basis of other tests of oral syntactic proficiency. A number of tests are available for assessing the oral language proficiency of children; most of these include some measure of syntactic proficiency. If any of these tests have already been demonstrated to be more direct measures of proficiency, then a statistically significant correlation between scores obtained on the BSM and a more direct measure would constitute concurrent validation. The ideal way to establish concurrent validity, however, would be to compare scores obtained on the BSM with real life, naturalistic language data for the same individuals. Such a procedure would ensure that an indirect measure such as the BSM actually reflects real life language proficiency.

No major validation study has yet been performed on the BSM as a proficiency test. However, results from several studies from the second language acquisition research literature bear directly on the question of the validity of the BSM. One study relates to the first means of demonstrating concurrent validity, i.e. through comparison with other proficiency measures. A second study compares morpheme scores obtained using the BSM as an elicitation instrument with morpheme scores obtained from spontaneous speech data from the same individuals on the same day.

Larsen (1975a and b) examined the correlations between the accuracy order of morphemes obtained from data elicited by the BSM and from several other measures of language proficiency. The morpheme accuracy order from only one of her measures, an elicited imitation task, attained a significant correlation (rho=.72; $p<.05$) with the order Dulay and Burt (1974b) report. However, Larsen's measures, like the BSM, were not subjected to validation.

Another study which is suggestive, although inconclusive due to the small sample, is an examination of both spontaneous speech data and data elicited by the BSM (Rosansky, 1976a and b; 1977). This was done as part of an investigation into second language acquisition research. In a direct comparison of BSM-generated and spontaneous speech-based orders of morphemes, for two subjects, the ranked percent correct orders of morphemes derived from spontaneous speech data and from BSM-elicited speech did not correlate for the same person at the same point in time. Moreover, the two data collection techniques did not yield the same correlation (BSM: .68; spontaneous: -.50) in a cross-comparison of two subjects' orders of morphemes compared with each other. In these cases, at least, Rosansky suggests that it seems that the BSM "does not validly measure the same thing as spontaneous speech data, and that this may be due to artifacts within the test itself (1976b, p. 95)."

Item Analysis

Items on a test can be analyzed both qualitatively, i.e. with regard to their content and form, and quantitatively, i.e. with regard to their statistical characteristics. A qualitative analysis often entails the consideration of content validity and of the way in which items are written, while a quantitative analysis entails the measurement of item difficulty and item validity. "Both the validity and the reliability of any test depend ultimately on the characteristics of its items (Anastasi, 1976, p. 198)."

An examination of the content validity of the BSM, particularly with regard to the range of questions designated as Level 5 items, appears in the following section. Sanchez (1976) discusses several points which can be considered in conjunction with a qualitative analysis of the items on the BSM. She cites several dialect problems due to morphological variants within Spanish that reflect on the BSM-S. In addition, she observes that certain of the items on the BSM-S are "unnatural" and in some cases "they violate selective feature constraints (p. 124)."

A quantitative analysis of test items includes the consideration of item validity. Item validity is essentially a statistical measure of the relationship between item responses and overall criterion performance, or between sub-scales and overall performance. The BSM is a test of only 18 items, i.e. the child's syntactic ability will be assessed on the basis of correct-incorrect responses to 18 items. While these 18 items have been carefully selected and further divided into Level 5 or Level 3-4 items on the basis of extensive field testing, there is some uneasiness about basing a child's syntactic proficiency level, and thus the assessment of degree of dominance or mastery of the language, on so small a sample of utterances.

The two sets of items (i.e. Level 5 and Level 3-4) were carefully analyzed during the test development, and the authors report point biserial correlations. The point biserial correlations for Level 5 items on the BSM-E range from .08-.48, and from .21-.84 for the BSM-S. Correlations for the Level 3-4 items were in the .66 range on the BSM-E (with one item reaching .92), and in the .85 range on the BSM-S. Items with very low correlations were discarded from the test, yet many of the correlations (after Henrysson's correction procedure for spuriousness arising from including an item in the total score on a short scale) were still quite low. Burt et al. point out the low correlations in their discussion, and suggest that these values could have been deflated due to extremely easy and extremely difficult items. This explanation is surprising, since "[The point biserial correlation coefficient] yields a measure of item-criterion relationship that is independent of item difficulty (Anastasi, 1976, p. 214)." Furthermore, this measure of item validity "assumes a continuous and normal distribution of the trait underlying the dichotomous item response (p. 214)," and it is not clear whether the data do, in fact, meet this assumption.

Content Validity

Content validity is evaluated by determining to what extent the content of a test (i.e. the range of items) actually samples the kinds of situations or the subject matter about which conclusions will be drawn. Content validity, like construct validity, is not a statistically calculable type of validity. Nevertheless, it is important to demonstrate this type of validity (Peterson and Cartier, 1975).

Close examination of the items on the BSM raises some serious questions about content validity. While the BSM is a test of syntax, and all of the items are scored for "grammatical correctness," there are some practical problems which arise with regard to the eight Level 5 items. Since it is only these eight items which presumably discriminate between assignment to Level 5 or Level 3-4, we must be sure that these eight items do, in fact, serve this function.

Among the Level 5 items, there are essentially two places where the production of each of four structures (possessive 's, long plural -es, past irregular, and conditional perfect) are possible. To be assigned to Level 5 proficiency, the child need only produce six of these eight items correctly. Two of the test items were designed to elicit conditional perfect, e.g.

What would have happened to his shoes if he hadn't taken them off?

Only 6 percent of the field tested children actually scored correctly on the two conditional perfect items. Thus, children placed in Level 5 can do so with a score of 6/8, and they need not demonstrate mastery of the conditional perfect.

One problem with the items in Level 5 is that not all items are equally good at forcing the desired structures in the children's responses. While the two items (Nos. 13 and 23) designed to elicit the conditional perfect do ensure that the responses must contain the conditional perfect, e.g.

(in response to Item No. 13)
They would've gotten wet,

this is not true for those items eliciting the possessive 's (Items No. 18 and 20), e.g.

Item No. 20
Whose flower is that?

In this example, the child could equally answer with the possessive pronoun hers as with the girl's. Since all responses which are grammatically correct are scored as correct, even if the desired structure is not produced, the response hers in answer to Item No. 20, would be scored correct. The same is true with Item No. 18. Thus, the child could be scored correct on these two items without demonstrating mastery of the possessive 's morpheme.

The two long plural items, like the two items eliciting conditional perfect, both seem to force a long plural -es response, e.g.

> (In Items No. 6 and 7, the examiner points to two houses in one picture, and later to two noses on the houses, and inquires:)
> What are these?

On the other hand, the items designed to elicit past irregular morphemes could also elicit other grammatically correct responses, e.g.

> Item No. 22
> What happened to the King's food?
>
> *Hoped for response:*
> The dog ate it.
>
> *Other possible responses:*
> It was eaten.
> It's gone.
>
> Item No. 24: (pointing to a picture of an apple on the floor)
> What happened to that apple?
>
> *Hoped for response:*
> It fell.
>
> *Other possible responses:*
> Someone dropped it.
> It rolled off.

Thus, in fact, a child could be placed in Level 5 proficiency (i.e. could score 6/8 on the eight Level 5 items) by missing the two conditional perfect items, and, theoretically, not actually demonstrating mastery of either possessive 's or the past irregular. In a sense, then, the only discriminating items for Level 5 might be the two items eliciting long plural (i.e. the two responses houses and noses).

The long plural structure is actually only partly a test of syntactic knowledge. The syntactic knowledge of plural (i.e. knowing in which linguistic environments to mark an English noun with one of the plural allomorphs (/-s, -z, ɪz/) is equally demonstrated by the short plural morpheme /-s/. The short plural is considered to be a lower level structure by Burt et al.; nearly 84 percent of their field tested subjects supplied this structure correctly. The /ɪz/allomorph or long plural, in fact, tests for only additional knowledge of an English *phonological* rule; *syntactic* knowledge of this structure is exhibited in the lower level structure.

According to Burt et al., the BSM is a test of syntactic proficiency. Yet, in the final form of the test, it would be theoretically possible (as demonstrated above) for the student to be placed in Level 5 by failing to provide the hoped for responses to the two conditional perfect items and by giving grammatically correct but lower level structure answers for the two past tense irregular and the two possessive items. The use of these lower level structures (i.e. houses, noses, etc.) depends on English phonological, or at the best morphophonemic, rules rather than on syntactic knowledge. If, then, the only unambiguously discriminating item between the top level of the test and the other two depends (at least in part) on phonological knowledge, rather than on syntactic knowledge, the utility of the test and what it discriminates becomes questionable.

Finally, since the BSM is designed to measure syntactic proficiency, it is important to point out that the test does not, in reality, measure a range of syntactic abilities which one might consider to be within the realm of syntactic proficiency. For example, the authors note that because of the format and structure of the test, it is not possible to elicit such syntactic structures as interrogatives, passives, demonstratives, and relative clauses. The test is, in fact, limited to an assessment of a small subsample of morphemes, and does not sample the broad range of syntactic proficiency.

CONCLUSION

I have discussed the advantages and shortcomings of the Bilingual Syntax Measure both as a test of oral language proficiency and as a research tool. Potential users will have to evaluate the usefulness of the BSM by weighing its ease of administration and attractiveness against the test's serious technical weaknesses of low reliability and questionable validity. A major problem still remains for researchers in the field and for test users in the schools. I encounter this problem frequently. Students needing to collect language data for term projects and school personnel wishing to assess the English or Spanish proficiency of their students come to me and ask for my evaluation of the BSM. When I point out the deficiencies outlined here, the response is always the same: "What, then, is a good, easily administered test to use for the assessment of second language proficiency of children?"

Levine (1976) has considered the problem of testing communicative competence. She notes that language testing has "dealt almost exclusively with the testing of those skills which, separately, are said to make up language....Such tests are widely used because they are technically objective, and...they are relatively simple to administer and to mark (p. 128)." However, she points out that "because of the desire to test only those aspects which can be validly and reliably tested, the full range of testing techniques is not used and the full range of skills is not tested (p. 128)." Drawing on Cooper's (1968) testing model, she suggests language abilities can be divided into two aspects--content (e.g., phonological, grammatical, semantic, etc.) and knowledge of contextual restraints--as well as into four modes (e.g., aural comprehension, oral production, reading, and writing). Based on these two aspects and four modes, Cooper developed a matrix containing roughly 20 different language skills. A consideration of the BSM in terms of Cooper's matrix reveals that of the 20 language skills, the BSM taps only four: oral production and aural comprehension of limited morphology and limited syntax. Thus, while the BSM is designed to assess oral/aural syntactic abilities, it does so only to a limited extent.

Rudd (1969) in referring to the language testing of immigrant children in Britain observes that, "For tests to be really relevant to children's needs, we should also...test their ability to use...language items/skills *in* relevant activities. It may not prove possible to do this validly and reliably, but if we want to know about the children's ability to *use* English in practice, which is what counts in the long run, then we must try." This indeed is the crux of our problem. To my knowledge, we have no such instrument. With the ever-increasing demands from schools and the continued interest of researchers in bilingual education, we must turn to the development of an adequate measurement instrument as a top priority. Continued use of the BSM as a research tool, with its questionable validity, leaves many otherwise well-designed research studies subject to serious criticism at their very core--the data base. Continued placement of children in bilingual programs on the basis of the BSM, with its limited sample of items and fuzzy proficiency level assignments, is also risky.

What we need now is a major effort at evolving a test that is both easy to administer and technically excellent, designed for assessing the second language proficiency of children. I am hopeful that with new Federal funds being made available for research in bilingualism and bilingual education, such a major effort will soon take place.

FOOTNOTES

[1] There are 14 structures referred to here (although elsewhere calculations are done on only 11). From Table 3 of the *Technical Manual*, the structures examined were: word order, pronoun case (scored for 3rd person singular and plural in subject position), progressive -ing, short plural -s, long plural -es, auxiliary (3rd person be), copula (3rd person), article (definite and indefinite is same category), 3rd person present tense singular (for regular verbs), past regular, past irregular, possessive 's, and conditional perfect (would + have + -en). The BSM-S structures included word order, progressive (present), auxiliary (3rd person singular present form of estar, copula (3rd person singular present tense of ser and estar), 3rd person singular present, infinitive (scored together for all -ar, -er, and -ir forms), adjective gender, possessive de and del, conjunction (que), reflexive se in verb-particle forms and independently as in "se la comio," presence of article (and gender, if present), present subjunctive (for 3rd person regular and irregular forms), direct and indirect object pronouns, and past subjunctive (perfect) and past participle.

[2] These include long plural, perfect conditional, possessive, past irregular, short plural, copula, auxiliary, article, and progressive -ing. The BSM-S was designed to elicit the following structures on the final version: present and past (perfect) subjunctive, reflexive (se), direct and indirect object pronouns, conjunction (que), present indicative, possessive, article, adjective gender, copula, progressive, and auxiliary.

[3] As was demonstrated on page 120.

[4] My own experience administering the BSM to Spanish speaking subjects acquiring English as a second language (see Cazden et al., 1975) was quite similar. We fortunately had recourse to tape recordings of the sessions.

Bibliography

Anastasi, A. *Psychological Testing*. 4th ed. New York: Macmillan, 1976

Angoff, W.H. "Scales, Norms, and Equivalent Scores." In R.L. Thorndike, ed., *Educational Measurement*. 2nd ed. Washington, D.C.: American Council on Education, 1971, pp. 508-600

______ and A.T. Sharon. "A Comparison of Scores Earned on the Test of English as a Foreign Language by Native American College Students and Foreign Applicants to U.S. Colleges." *TESOL Quarterly* 5:129-136, 1971

Bart, W.M. and D.J. Krus. "An Ordering Theoretic Method to Determine Hierarchies among Items." *Educational and Psychological Measurement* 33:291-300, 1973

Bailey, N., C. Madden, and S.D. Krashen. "Is There a 'Natural Sequence' in Adult Second Language Learning?" *Language Learning* 24:235-243, 1974

Brown, R. *A First Language: The Early Stages*. Cambridge, Mass.: Harvard University Press, 1973

______, C.B. Cazden, and U. Bellugi. "The Child's Grammar from I to III." In J.P. Hill, ed., *1967 Minnesota Symposium on Child Psychology*. Minneapolis: University of Minnesota Press, 1969

Burt, M.K., H.C. Dulay, and E. Hernandez-Chavez. *Bilingual Syntax Measure*. New York: Harcourt Brace Jovanovich, 1975

Carroll, J.B. "Fundamental Considerations in Testing for English Language Proficiency of Foreign Students." In *Testing the English Proficiency of Foreign Students*. Washington, D.C.: Center for Applied Linguistics, 1961, pp. 30-40

______. "The Psychology of Language Testing." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach*. London: Oxford University Press, 1968, pp. 49-69

______. "Note on the Scoring of Foreign Language Speaking and Writing Fluency Tests." *Educational Testing Service Research Bulletin*, RB-70-52, September 1970 [ERIC Document Reproduction Service Document ED 044 441]

______. "Foreign Language Testing: Will the Persistent Problems Persist?" In M.C. O'Brien, ed., *Testing in Second Language Teaching: New Dimensions*. Dublin: University Press, 1973, pp. 6-17

______. *The Teaching of French as a Foreign Language in Eight Countries*. (International Studies in Evaluation, V.) Stockholm: Almqvist & Wiksell and New York: John Wiley & Sons, 1975

Cazden, C.B., H. Cancino, E.J. Rosansky, and J.H. Schumann. "Second Language Acquisition Sequences in Children, Adolescents and Adults." Final Report, Project No. 730744, Grant No. NE-6-00-3-0014, U.S. Department of Health, Education and Welfare, National Institute of Education, 1975

Clark, J.L.D. *Foreign Language Testing: Theory and Practice*. Philadelphia: Center for Curriculum Development, 1972a

______. "Measurement Implications of Recent Trends in Foreign Language Teaching." In D.L. Lange, ed., *Foreign Language Education: A Reappraisal*. (ACTFL Review Series, 4.) Skokie, Ill.: National Textbook Co., 1972b, pp. 219-257.

______. "Theoretical and Technical Considerations in Oral Proficiency Testing." In Jones and Spolsky, eds., 1975, pp. 10-28

______. *The Performance of Native Speakers of English on the Test of English as a Foreign Language*. (TOEFL Research Reports, 1.) Princeton, N.J.: Educational Testing Service, 1977

Cohen, J. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20:37-46, 1960

_______. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70:213-220, 1968

Cronbach, L.J. "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika* 16:297-334, 1951

Darnell, D.K. *The Development of an English Language Proficiency Test of Foreign Students Using a Clozentrophy Procedure.* Boulder, Co.: Department of Speech and Drama, University of Colorado, 1968

Davies, A.L. "Introduction." In J.P. Allen and A. Davies, eds., *The Edinburgh Course in Applied Linguistics: Testing and Experimental Methods.* (Language and Language Learning Series, 4.) London: Oxford University Press, 1977a

_______. "The Construction of Language Tests." In J.P. Allen and A. Davies, eds., 1977b

de Villiers, J. and P. de Villiers. "A Cross-sectional Study of the Development of Grammatical Morphemes in Child Speech." *Journal of Psycholinguistic Research* 2:267-278, 1973

Dulay, H.C. "Aspects of Child Second Language Acquisition." Unpublished PhD dissertation, Harvard University, 1974

_______ and M.K. Burt. "Should We Teach Children Syntax?" *Language Learning* 23:245-258, 1973

_______ and M.K. Burt. "Errors and Strategies in Child Second Language Acquisition." *TESOL Quarterly* 8:129-136, 1974a

_______ and M.K. Burt. "Natural Sequences in Child Second Language Acquisition." *Language Learning* 24:37-53, 1974b

Educational Testing Service. *Test of English as a Foreign Language: Interpretive Information.* Princeton, N.J.: Educational Testing Service, 1967

_______. *Test of English as a Foreign Language: Interpretive Information.* Princeton, N.J.: Educational Testing Service, 1968

_______. *Test of English as a Foreign Language: Interpretive Information.* Princeton, N.J.: Educational Testing Service, 1970

_______ and College Entrance Examination Board. *Manual for TOEFL Score Recipients.* Princeton, N.J.: Educational Testing Service and New York: College Entrance Examination Board, 1973

Goldberg, L.R. "Man versus Model of Man: A Rationale, Plus Some Evidence for a Method of Improving on Clinical Inferences." *Psychological Bulletin* 73:422-432, 1970

Horst, P. "Correcting the Kuder-Richardson Reliability for Dispersion of Item Difficulties." *Psychological Bulletin* 50:371-374, 1953

Hunt, K. "An Elementary School Curriculum to Develop Better Writing Skills." Final Report, U.S. Department of Health, Education and Welfare. Mimeo. Tallahassee: Florida State University, 1970

Husen, T., ed. *International Study of Achievement in Mathematics: A Comparison of Twelve Countries.* 2 vols. New York: John Wiley & Sons, 1967

Irvine, P., P. Atai, and J.W. Oller, Jr. "Cloze, Dictation, and the Test of English as a Foreign Language." *Language Learning* 24:245-252, 1974

Jameson, S.C. and S.J. Malcolm. "TOEFL--The Developing Years." *International Cultural and Educational Exchange,* 57-62, Winter 1972-73

Jones, R.L. "Testing Language Proficiency in the United States Government." In Jones and Spolsky, eds., 1975, pp. 1-9

_______. "Achieving Objectivity in Subjective Language Tests." *Proceedings of the Fourth International Congress of Applied Linguistics.* Stuttgart: HochschulVerlag, 1976, pp. 241-253

_______ and B. Spolsky, eds. *Testing Language Proficiency.* Arlington, Va.: Center for Applied Linguistics, 1975

Kaiser, H.F. "The Varimax Criterion for Analytic Rotation in Factor Analysis." *Psychometrika* 23:187-200, 1958

Kaplan, R.B. "TOEFL in 1969--A Reappraisal." *NAFSA Newsletter* 21, No. 1:7-8, 10, September 1969

Krashen, S.D., C. Madden, and N. Bailey. "Theoretical Aspects of Grammatical Sequencing." In M.K. Burt and H.C. Dulay, eds., *New Directions in Second Language Learning, Teaching, and Bilingual Education.* Washington, D.C.: Teachers of English to Speakers of Other Languages, 1975

Kuder, G.F. and M.W. Richardson. "The Theory of the Estimation of Test Reliability." *Psychometrika* 2:151-160, 1937

Lado, R. *Language Testing*. London: Longmans, Green and Co., 1961

Larsen, D. "The Effect of Formal Language Instruction on the Order of Morpheme Acquisition." Paper presented at the 9th Annual Convention of Teachers of English to Speakers of Other Languages, Los Angeles, March 4-9, 1975a

_______. "The Acquisition of Grammatical Morphemes by Adult Learners of English as a Second Language." Unpublished PhD dissertation, University of Michigan, 1975b

Lebach, S. Interim report from the Culver City Spanish Immersion Project. Unpublished (and untitled) report, UCLA, 1975

Levine, J. "An Outline Proposal for Testing Communicative Competence." *English Language Teaching Journal* 30:128-135, 1976

Levine, R.S. "Equating the Score Scales of Alternate Forms Administered to Samples of Different Ability." *"Educational Testing Service Research Bulletin,* RB-55-23, November 1955

Lewis, E.G. *Linguistics and Second Language Pedagogy*. The Hague: Mouton, 1974

Lord, F.M. and M.R. Novick. *Statistical Theories of Mental Test Scores*. Reading, Mass.: Addison-Wesley, 1968

Lowe, P., Jr. "The Oral Interview--A Criterion-Referenced Test?" Paper presented at the annual meeting of the American Council on the Teaching of Foreign Languages, November 1975 [ERIC Document Reproduction Service Document ED 119 473]

_______. "Handbook on Question Types and Their Use in LLC Oral Proficiency Tests." Mimeo. Beirut: American University of Beirut, 1976

Lyman, H.B. *Test Scores and What They Mean*. 3rd ed. Englewood Cliffs, N.J.: Prentice-Hall 1978

National Council of Teachers of English. "Report of the Meetings of the Test Planning and Test Writing Committees of the National Council on the Testing of English as a Foreign Language." Mimeo. Urbana, Ill.: NCTE, 1963

Osgood, C.E., J. Suci, and P.H. Tannenbaum. *The Measurement of Meaning*. Urbana, Ill.: University of Illinois, 1957

Perren, G.E. "Testing Spoken Language: Some Unsolved Problems." In A. Davies, ed., *Language Testing Symposium: A Psycholinguistic Approach*. London: Oxford University Press 1968, pp. 107-116

Petersen, C.R. and F.A. Cartier. "Some Theoretical Problems and Practical Solutions in Proficiency Test Validity.: In Jones and Spolsky, eds., 1975, pp. 105-118

Pike, L. "An Evaluation of Present and Alternative Item Formats for Use in the Test of English as a Foreign Language." Mimeo. Princeton, N.J.: Educational Testing Service, 1973

Porter, J. "A Cross-sectional Study of Morpheme Acquisition in First Language Learners." Unpublished paper, Harvard University, 1975

Rand, E. "Integrative and Discrete Point Tests at UCLA." *UCLA Workpapers in TESL 6:87-92, 1972*

Rice, F.A. "The Foreign Service Institute Tests Language Proficiency." *The Linguistic Reporter* 1, No. 2:4, 1959

Rosansky, E.J. "Second Language Acquisition Research: A Question of Methods." Unpublished PhD dissertation, Harvard University, 1976a

_______. "Methods and Morphemes in Second Language Acquisition Research." *Language Learning* 26:409-425, 1976b

_______. "Explaining Morpheme Accuracy Orders: Focus on Frequency." In C. Henning, ed., *Proceedings of the Los Angeles Second Language Research Forum*. Los Angeles: UCLA, 1977

Rudd, E. "Language Tests--What Do We Want Them to Do?" *English for Immigrants* 3, No. 1, 1969

Sanchez, R. "Critique of Oral Language Assessment Instruments." *NABE Journal* 1:120-127, 1976

Silverman, R.J., J.K. Noa, and R.H. Russell, eds. *Oral Language Tests for Bilingual Students: An Evaluation of Language Dominance and Proficiency Instruments*. Portland, Or.: Northwest Regional Educational Laboratory, 1976

Slocum, J.B. "TOEFL in 1969--An Appraisal." *NAFSA Newsletter 21*, No. 1:6-7, September 1969

Spolsky, B. "Do They Know Enough English: Some Notes on the Problems of Assessing the Proficiency in English of Foreign Students." In *NAFSA Studies and Papers*. (English Language Series, 13, ATESL Selected Conference Papers.) Los Altos, Calif.: Language Research Associates Press, 1967

_______, P. Murphy, W. Holm, and A. Ferrel. "Three Functional Tests of Oral Proficiency." *TESOL Quarterly* 6:221-235, 1972

Thorndike, R.L. and E.P. Hagen. *Measurement and Evaluation in Psychology and Education*. 4th ed. New York: John Wiley & Sons, 1977

Upshur, J.A. "Productive Communication Testing: Progress Report." In J.W. Oller, Jr. and J.C. Richards, eds., *Focus on the Learner: Pragmatic Perspectives for the Language Teacher*. Rowley, Mass.: Newbury House, 1973, pp. 177-183

Valette, R.M. *Modern Language Testing*. New York: Harcourt Brace Jovanovich, 1977

Walker, D.A. *The IEA Six Subject Survey: An Empirical Study of Education in Twenty-One Countries*. (International Studies in Evaluation, IX). Stockholm: Almqvist & Wiksell and New York: John Wiley & Sons, 1976

Wilds, C. "The Oral Interview Test." In Jones and Spolsky, eds., 1975, pp. 29-44

Winkler, R.L. "The Quantification of Judgement: Some Experimental Results." *Proceedings of the American Statistical Association* 62:386-395, 1967

Papers in Applied Linguistics

ADVANCES IN LANGUAGE TESTING SERIES: 1

---

1. Some Major Tests
   edited by Bernard Spolsky
2. Approaches to Language Testing
   edited by Bernard Spolsky

Bernard Spolsky, General Series Editor